PhantomDB: Using Generative AI To Spawn an Open-source Extensible “Phantom” Patient Database for AI/ML Research in AD/ADRD

Executive Summary

Understanding the Challenge: The PREPARE Challenge aims to revolutionize the early prediction of Alzheimer’s Disease and Alzheimer’s Disease Related Dementias (AD/ADRD). It seeks innovative datasets that can empower machine learning (ML) and artificial intelligence (AI) methodologies to predict AD/ADRD earlier than current clinical practices, with key emphasis on overcoming biases in existing datasets and clinical practices. AI algorithms are typically data hungry, requiring large datasets to train and validate upon. This results in a bottleneck since clinical databases are seldom open-sourced, partly due to privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) considerations in the US, and often due to the commercial and profit incentives of data warehouses which often have vast amounts of relevant data, but are unwilling to allow open source access[1]. Without open source access, profit motives stifle scientific progress. For example, the Marketscan Database of commercial insurance claims, that houses Electronic Healthcare Records (EHR) with emphasis on diagnostic and procedural records for over 100 Million US patients, costs upward of 100K USD per year to access, putting it squarely out of reach of a large section of the research community[2].

Our Dataset: “Phantom” Patients Brought to Life with Novel Generative AI: Our current and ongoing work is developing a novel generative AI framework that can produce medical histories reflecting realistic deep and often uncharted comorbidity constraints and relationships learned from an actual database of $>$20K 60-75year old patients eventually diagnosed with AD/ADRD. Our generative AI uses a novel approach to learn and characterize cross-dependencies between time-stamped diagnostic and procedural codes in individual patient histories, to identify a 774,627 parameter model. This model, the “phantom-net”, can replicate perturbed variations of medical histories from scratch, that are no longer proprietary data, but replicate delicate cross-talk across the entire human disease-spectrum for AD/ADRD patients.

We are sharing diganostic history of 2,000,000 phantom patients (1M phantoms generated from a national model, and 1M generated from Chicagoland African Americans treated at the University of Chicago Medical Center (UCM)), comprising time-stamped diagnostic and procedural codes, along with the generative model and software (python package teomim on PyPi), all with a permissive license. We will share the generative algorithm and results that demonstrate via critical stress-test comparisons that it is exceedingly difficult to differentiate between the training database and generated phantom patients, or between validation held-back patient samples and the phantom patients. The research community will now have access to a high quality EHR database, validated to closely replicate real data without licensing constraints.

Predictor Variables: PhantomDB comprises phantom patient diagnostic procedural codes in ICD10[3,4] format, timestamped with the age of the patient at the time these codes are “generated”. Code density, and timeline over which individual patients are visible in the PhantomDB replicate statistical characteristics of actual data.

Target AD/ADRD Variables: The target variables for AD/ADRD for the PhantomDB are diagnostic codes for AD/ADRD. Codes for mild cognitive impairment (MCI) or other code combinations that depict typical diagnostic pathways maybe used to construct robust target identifiers.

As a solution to the throttling of progress from the difficulty of accessing high-quality training data, our submission is arguably an exemplary candidate for the PREPARE Challenge, poised to significantly impact the early diagnosis and treatment strategies for AD/ADRD.

Data Description

Thus, the PREPARE Challenge envisions significant progress in the field of AD/ADRD research with its core objective being to revolutionize early detection methods by leveraging machine learning (ML) and artificial intelligence (AI). This can be a substantially improve patient outcomes because early detection has being demonstrated to improve disease management[5] by allowing for more effective treatment and planning, and providing a chance for patients to participate in emerging clinical trials. Also of crucial importance, this challenge aims to eliminate biases in current datasets and clinical procedures, seeking novel datasets that could facilitate AI and ML in predicting AD/ADRD earlier than current clinical practice.

The key issue that our submission aims to address is the barrier to progress arising from the reliance of AI algorithms on extensive training datasets. The strict regulatory environment in the context of healthcare hinders development, as necessary datasets can be difficult to access and utilize legally and ethically[6]. Furthermore, as discussed in commentary from Reuters (https://www.reuters.com/legal/litigation/data-privacy-artificial-intelligence-health-care-2022-03-17/), the use of AI in healthcare necessitates careful consideration of security issues. De-identifying patient health information before uploading it into an AI database is one way to navigate these regulations. However, the process of de-identification itself can be challenging, and as AI systems evolve, they may inadvertently re-identify data that was previously de-identified. This challenge is compounded by the need to continually assess the privacy risks as AI systems become more sophisticated and capable of creating data linkages that didn't previously exist.

This present data submission from our team proposes a novel approach towards alleviating these challenges. Our solution is to leverage the power of generative AI to produce data that would be indistinguishable from real patient populations, but belong to no actial human, freeing such data from regulatory control, and proprietary restrictions. Our submission comprises a large number of "phantom" patients with high quality timestamped diagnostic medical history, over a continuous time-span of 15 years, from age 60 to 75 of the phantom patients.

TABLE 1: Dataset Information

Number of Patients in MarketScan Proprietary Database	21,374 (National)
Number of Patients in UCM Database	187 (African-American from Chicagoland)
Phantom patients generated	1M (National)
African-American phantom patients generated	1M (Chicagoland)
Timespan	15 years of diagnostic history (age 60 - 75).
Unique ICD10 codes in PhantomDB	816 3-character ICD-10 prefixes
ICD10 codes per patient per year	National: 17.4, African American: 10.0

Basic Information: Electronic Health Record (EHR) diagnostic data, which primarily consists of timestamped diagnostic codes, usually follows the ICD-10 (International Classification of Diseases, Tenth Revision) format. While these data were originally used for administrative purposes like billing and insurance, they have increasingly become a valuable resource for studying and modeling the human disease spectrum. This is due to their comprehensive recording of various time scales, comorbidities, and disease prevalence dynamics across large databases of EHR records. Amongst the vast number of reserach papers that have begun to use such clinical or administrative data to design screening and early diagnostic tools for diseases ranging from Idiopathic Pulmonary Fibrosis[7] to Autism Spectrum Disorder[8] to cardiac risk after elective surgeries[9]. However EHR data, while offering exciting research opportunities, come with inherent challenges. These include issues related to data quality, the accuracy of recorded information, and the potential biases inherent in data collected primarily for clinical rather than research purposes.

To address these challenges, we have developed a novel generative AI framework capable of producing medical histories that reflect the complex and often unexplored comorbidity constraints and relationships learned from a database of 21,734 patients aged 60-75 (and 187 African-American patients treated at UCM), who are diagnosed with AD/ADRD as evidenced by their EHR record at some point. This generative AI, termed ”phantom-net ”, utilizes a 774,627 parameter model to learn and characterize cross-dependencies between time-stamped diagnostic codes in individual patient histories. This model can replicate perturbed variations of medical histories that are no longer proprietary data but simulate the intricate interactions across the human disease spectrum for AD/ADRD patients. Details of the dataset submitted, and size of the proprietary patient databases used to train the phantom-net is shown in Table 1.

Utility & Rigor: The utility of a large database of open source clinical EHR data, where the data sparsity is realistic but each patient is “visible” over 15 years, and captures known and unknown comorbidities cannot be overstated. Our validation results included show that the phantom database captures realistic medical history patterns, and thus may be used as an invaluable training and validation resource in the public domain.

phantom-net Construction: We describe the details of phantom-net construction and inference briefly. The phantom-net is a generative model that aims to capture the cross-dependecies between the occurrences of any the 816 ICD10 prefixes in an individual's medical history. Internally, for the purpose fo training, the medical history of individual patients from proprietary databses is represented in a tabular format, with columns corresponding to $\langle \mathrm{ICDCODE\_AGE} \rangle$, and rows corresponding to individual patients. To make sure we can accommodate maximum details of patient diagnostic history without loss, we encode column with three letter prefixes of ICD10 codes, and the remaining suffix for individual patients is actually entered as table entries. The AGE suffix in the column specification designates the age (in six month increments, first half denoted by a suffix 0, and second half of the age-year indicated by a 1, see Eq. (1)) at which the ICD10 prefix was encountered in the patient history. An example excerpt of the patent data encoding from this scheme can be described as follows:

Suppose `patient_A` has codes G72.9 and G24.5 at ages 61 years 3 months and 62 years 5 months and patient `patient_B` has codes G81.9 at age 62 years 7months, then a portion of the encoding dataframe will look like:

$$\left[\begin{array}{lcccccc} & \mathrm{G72\_61\_1} & \mathrm{G72\_61\_2} & \mathrm{G24\_62\_1} & \mathrm{G24\_62\_2} & \mathrm{G81\_62\_1} & \mathrm{G81\_62\_2} \\ \text{patient\_A} & 9 & - & 5 & - & - & - \\ \text{patient\_B} & - & - & - & - & - & 9 \end{array}\right] \tag{1}$$

The phantom-net then infers a generative model to capture the cross-dependencies of a priori unspecified complexity between the different columns of the tabular representation of the diagnostic code-histories of a patient database (in this case proprietary databases from MarketScan and UCM), all satisfying some phenotype of interest (which in this case is the presense of one or more codes for AD/ADRD). A successful inference will capture known and uncharted co-morbidities, and replicate any longitudinal effects that might exist. Formally, we have:

Definition 1 (phantom-net)**.** *Let $X \sim P$ be an n-dimensional discrete random vector supported on a finite set Σ and following distribution P, i.e.*

$$X = (X_1, \dots, X_n) \sim P, \qquad \mathrm{supp}(X) = \Sigma = \prod_{i=1}^{n} \Sigma_i \text{ with } |\Sigma| < \infty \tag{2}$$

For $i = 1, \dots, n$, let $P_i := P(X_i \,|\, X_j = x_j \text{ for } j \neq i)$ denote the conditional distribution of X_i given the values of the other components of X. Finally, for each $i = 1, \dots, n$, let Φ_i^P denote an estimate of the distribution P_i. Then the set $\Phi^P := \{\Phi_i^P\}_{i=1}^n$ is called a phantom-net for the population P.

When P is clear from context, we may omit the superscript and simply write $\Phi = \{\Phi_i\}$ to denote the phantom-net. While the definition allows for arbitrary estimators Φ_i, here phantom-nets are computed using conditional inference trees[10] (a variant of classification and regression trees) to compute each Φ_i. Each phantom-net component Φ_i is computed independently, allowing a network structure (with possible cyclic dependencies) to form amongst these estimators.

Sampling Perturbations Around Real Patients: The phantom-net allows us to rigorously compute bounds on the probability of a perturbation of a medical history producing a "valid" phantom-patient, namely capturing the idea that random occurrences of diagnostic codes are not realistic, and deep long-ranging constraints exist that shape medical history over time. With an exponentially exploding number of possibilities in which a medical history over a large set of items can vary, it is computationally intractable to directly model all possible variations or their inter-related emergence rules; nevertheless, we can constrain the possibilities using the patterns distilled by the phantom-net.

Ultra-high-Dimensional Sampling to Generate Phantom Patients: From the phantom-net, we can infer approximations to the full conditional distributions. As the collection of full conditionals can be shown to uniquely determine the full joint distribution (by the Hammersley-Clifford theorem[11]), we can take the phantom-net approximations as a model of the joint distribution. Specified in this form, we obtain an efficient method of sampling the (high-dimensional) distribution without explicitly representing it. In particular, while it would be computationally difficult to directly sample a model distribution over hundreds or thousands of variables, here we can leverage the inferred conditionals in a natural way. In particular, starting from a known sample, we may iteratively update its indices by sampling the corresponding conditional distribution in the phantom-net. We then proceed to sample the next index, now using the generated value.

This procedure can be used both to generate new, realistic samples reflecting the model dependencies, as well as impute missing values that may be present in the data. To clarify, suppose the kth sample is $\mathbf{x}_k = (x_{k1}, \dots, x_{kn})$. We may define an indicator of missing values $\mathbf{m}_k = (m_{k1}, \dots, m_{kn})$ where $m_{kj} = 1$ if x_{kj} is missing and $m_{kj} = 0$ otherwise. The validity of this sampling routine has been proven theoretically in our recent work [REF], justifying the scheme:

1) Choose an index $j \in \{1, \dots, n\}$ for which $m_{kj} = 1$ and impute feature x_{kj} by sampling the distribution $\Phi_k(X_k | X_i = x_i, k \neq i)$
2) Go back to step 1, until no unobserved entity remains

Schematically, this procedure is similar to the well-known Gibbs sampling routine[12,13], which also uses iterative samples from full conditional distributions to generate samples from the joint distribution asymptotically. However, unlike Gibbs sampling q-sampling uses fixed approximate conditional distributions inferred by the phantom-net, and initializes from a known sample, which can allow q-sampling to converge faster.

Innovation: Our PhantomDB dataset is the first of its kind to the best of our knowledge to offer a large database of digital twins of proprietary clinical diagnostic histories. Additionally, we provide the generative AI model along with its trained parameters, so arbitrary new datasets can be spawned at will. The availability of such data in the public domain opens up new possibilities for training and testing ML algorithms. The key innovations may be enumerated as:

- **Data Accessibility and Ethics:** phantom-net circumvents data access barriers and aligns with ethical standards in digital healthcare by replicating medical histories without infringing on proprietary data rights, promoting a collaborative and inclusive research ecosystem.
- **Demographic-Specific Modeling:** Addressing disparities in AD/ADRD incidence, we focus on developing generative models for diverse demographics, starting with race-blind and African-American populations. This approach advances personalized healthcare and promotes equitable outcomes.
- **Technological Innovation with "Phantom" Patients:** Our generative AI framework, ca-

pable of producing realistic medical histories, marks a technological leap. By offering a high-fidelity replication of patient data free from proprietary constraints, it surpasses existing datasets in quality and applicability.

- **Extensive and Diverse Data:** The distribution of 2 million phantom patient histories, inclusive of a national model and a Chicago-specific African American model, greatly enhances the scale and diversity of data for AD/ADRD research, facilitating advanced machine learning applications.
- **Machine Learning Potential:** The provided dataset, validated to closely mimic real patient data, unlocks new potentials in machine learning for higher accuracy and earlier AD/ADRD predictions, heralding a new era in data-driven healthcare solutions.

Sample Characteristics and Representation: The dataset comprises diagnostic histories of 2,000,000 "phantom" patients, created using a novel generative AI framework. These digital twins of clinical diagnostic histories are obtained from 774,627 parameter generative AI (the phantom-net). Of these phantom patients, 1 million are generated from a national model reflecting a diverse population, while another 1 million are derived specifically from the African-American population in the Chicago area, treated at the University of Chicago Medical Center. Both patient cohorts on which the phantom-nets are trained belong to the age bracket 60-75 years, and are eventually diagnosed with AD/ADRD. This two model approach addresses the critical need for race and ethnicity-specific models in medical research, given the documented disparities in AD/ADRD incidence and care.

The national training cohort encompass a wide range of demographic characteristics such as age (60-75), race, sex, ethnicity, and other known demographics. The phantom patients are timestamped with diagnostic and procedural codes in the ICD10 format, reflecting the age of the patient at the time these codes are generated. The dataset not only captures the complexity of medical histories but also ensures representativeness and diversity, addressing the historical bias in AD/ADRD research.

The PhantomDB dataset, along with the generative model and software (available via the Python package 'teomim' on PyPi), is shared with a permissive license. Rigorous stress-tests validate the indistinguishability of phantom patients from real patient data, demonstrating the dataset's quality and reliability for research purposes.

Validation and Confidence in PhantomDB We show by a sequence of evaluation measures (Fig. 1) that the phantoms are not recognzable from human patients, either by a general classifier (panel a), code distribution (panel b), or disambiguation from control patients who dont have AD/ADRD diagnosis (panel c). It is also shown that PhantomDB has non-trivial correlation structure (panel d), are not correlated with the propritary training patients (panel e), and replicate prevalences of top Ad/ADRD comobidities (panel f). The prevalence of the top comorbidities are also shown in Table 2

Usability: The data is shared on Zenodo (https://doi.org/10.5281/zenodo.10598052). Zenodo objects have persistent digital object identifiers (DOI), which implies that this link will continue to be accessible forever. The data is distributed under **Creative Commons Attribution 4.0 International**, which allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately credited. The phantom patients have no identifiable information, and have very low or no correlation with training dataset, and thus cannot be re-identified with the training database.

The data is distributed as a compressed JSON object, which encodes time-stamped diagnostic codes for 2M phantom patients. The race of the patients is specified as a key in the json object. All patients are assumed to begin being observed at age 60. Full description of the format is given in the accompanying **data dictionary** (data_dictionary.pdf), and schema files

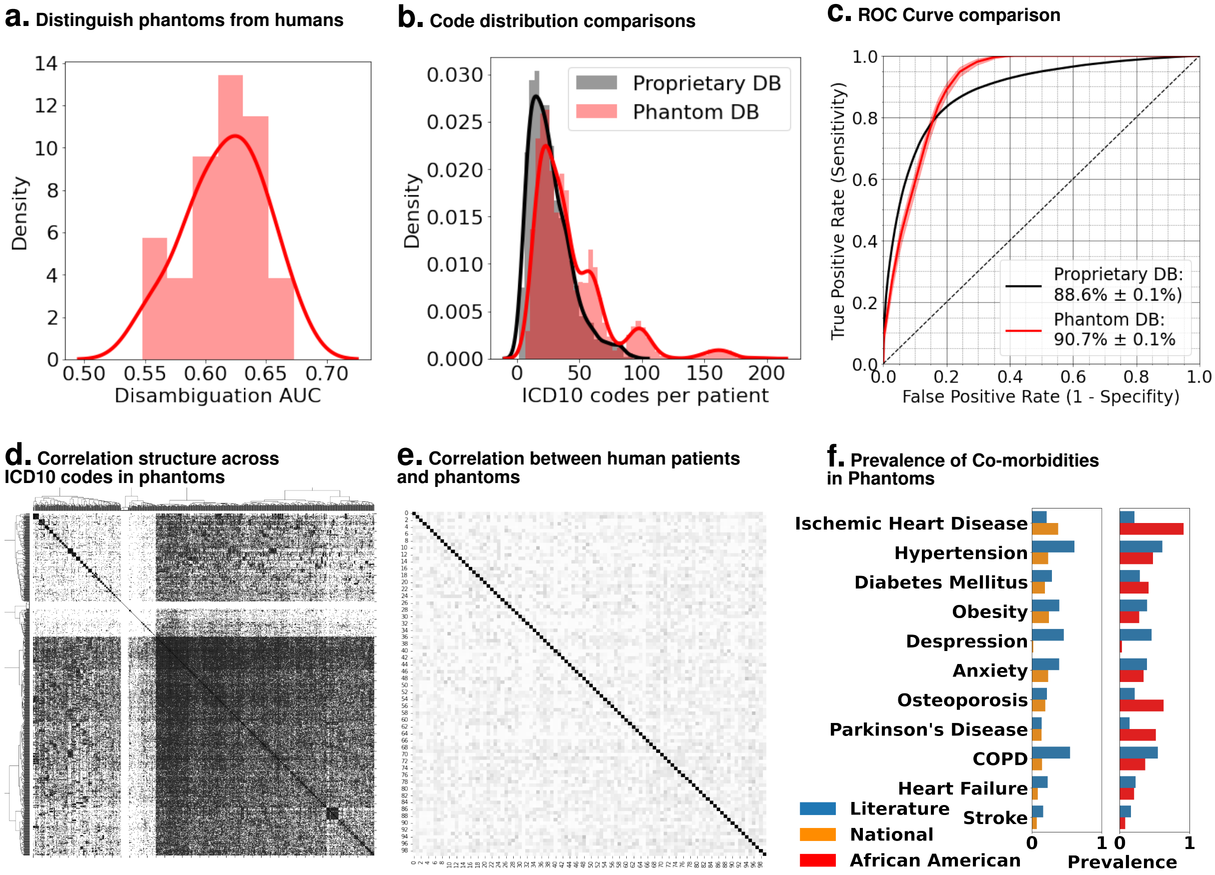

Fig. 1: **Evaluating Indistinguishability of Phantoms**. Panel a. Building a classifier to recognize phantoms from real patients based on the ICD10 code presence/absence achieves a mean AUC around 60%. The human patients are not observable over all 15 years, which helps the AUC to be $> 50\%$, but confirms that the phantoms are not recognizably different. Panel b. the distribution of the number of codes per patient is matched up closely between the human databases and the phantoms, despite being a non-trivial distribution. Panel c. Computing ROC curve to classify between human AD/ADRD positive patients from control human patients (who dont get diagnosed with dementia) achives comparable performance with the phantoms. Panel d shows the non-trivial correlation between co-morbidities captured in the PhantomDB, Panel e shows that there is no correlation between the human patients that inform the model and the phantoms, and panel f shows that the prevalence profile of top AD/ADRD comorbiditiesd match up relatively closely between literature and PhantomDB.

(schema.json). Additionally, all ICD10 prefixes that are used in the model are enumerated in the accompanying file *USED_ICD_CODES.xlsx*. These metadata files are also included in the Zenodo object version 1.1 (https://doi.org/10.5281/zenodo.10601248). To be more specific:

- The compressed PhantomDB database is named https://zenodo.org/records/10598052/files/phantomDB.tgz?download=1
- The phantom-net models for national and African-American cohorts are available at https://zenodo.org/records/10598052/files/national.pkl.gz?download=1 and https://zenodo.org/records/10598052/files/chicago_AA.pkl.gz?download=1.
- The data dictionary is at https://zenodo.org/records/10601248/files/data_dictionary.pdf?download=1
- The schema file is at https://zenodo.org/records/10601248/files/schema.json?download=1

Team Introduction: **Ishanu Chattopadhyay, PhD**, *Assistant Professor of Medicine at the University of Chicago*, is an expert in artificial intelligence, machine learning, and the computational aspects of data science. Chattopadhyay has been funded by the US Department of Defense (DARPA), the National Institute for Health, the Alzheimer's Association. Dr.

TABLE 2: Prevalences of Top AD/ADRD Co-morbidities (literature vs training vs phantoms)

Disease	Comorbidities in Literature	National Training Co-morbidities	UCM African-American Training Co-morbidities	National PhantomDB Co-morbidities	African-American PhantomDB Co-morbidities
Ischemic Heart Disease	0.211 [14–16]	0.53	0.30	0.38	0.91
Hypertension	0.609 [17]	0.90	0.17	0.23	0.48
Diabetes Mellitus	0.287 [17]	0.45	0.11	0.19	0.41
Obesity	0.392 [17]	0.14	0.09	0.24	0.28
Despression	0.456 [17]	0.39	0.01	0.02	0.03
Anxiety	0.39 [18]	0.31	0.59	0.23	0.34
Osteoporosis	0.214 [19,20]	0.31	0.13	0.19	0.63
Parkinson's Disease	0.14 [15,21,22]	0.09	0.21	0.14	0.52
COPD	0.546 [15,23,24]	0.37	0.08	0.14	0.36
Heart Failure	0.226 [15,25,26]	0.40	0.05	0.08	0.21
Stroke	0.161 [15,27,28]	0.27	0.06	0.07	0.08

Chattopadhyay won the prestigious Young Faculty Award from the Defense Advanced Research Projects Agency (DARPA) in 2020 for his work on formal methods to study cognitive dissonance and opinion dynamics. His current work on early screening algorithms fro AD/ADRD is funded by the Alzheimer's Association.

Dmytro Onishchenko, MS, *Research Staff Scientist* is an expert in software implementations of complex machine learning architectures, and leads the implementation of the PhantomDB project.

James A. Mastrianni, MD, PhD *Professor of Neurology, University of Chicago, Program Director, Behavioral Neurology, and Helen McLoraine Neuroscientist of the Brain Research Foundation.* Dr. Mastrianni is a Professor of Neurology, and serves as the Director of the memory Center at the University of Chicago. As part of a world-renowned academic medical center, The University of Chicago Medicine Memory Center offers leading-edge care by fully integrating the newest scientific research with the most current clinical knowledge. He and his team of neurologists, neuropsychologists, geriatricians, nurses, psychiatrists, and social workers provide comprehensive diagnostic evaluations and long-term management of patients with Alzheimer's disease and related dementias, especially atypical dementia and younger-onset Alzheimer's Disease. He also conducts clinical and basic science research aimed to better understand and treat neurodegenerative diseases.

Robert Gibbons, PhD *Blum-Riese Professor and a Pritzker Scholar at the University of Chicago,* is an esteemed figure in the field of statistics, holding appointments in the Departments of Medicine, Public Health Sciences, and Comparative Human Development. He is uniquely qualified to oversee the statistical rigor of our work, ensuring adherence to standard statistical guidelines in areas such as sampling and data analysis. As a distinguished member of several prestigious statistical organizations and the author of over 300 peer-reviewed scientific papers and five books, Professor Gibbons' expertise spans longitudinal data analysis, item response theory, environmental statistics, and drug safety.

References

[1] Asche, C. V., Seal, B., Kahler, K. H., Oehrlein, E. M. & Baumgartner, M. G. Evaluation of healthcare interventions and big data: review of associated data issues. *Pharmacoeconomics* **35**, 759–765 (2017).

[2] Reuters, T. Marketscan® commercial claims and encounters database. 2009 (2007).

[3] Organization, W. H. *ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision, 2nd ed* (World Health Organization, 2016).

[4] Manchikanti, L., Falco, F. J. & Hirsch, J. A. Ready or not! here comes icd-10. *Journal of neurointerventional surgery* (2011).

[5] Vinze, S., Chodosh, J., Lee, M., Wright, J. & Borson, S. The national public health response to alzheimer's disease and related dementias: Origins, evolution, and recommendations to improve early detection. *Alzheimer's & Dementia* **19**, 4276–4286 (2023).

[6] Hulsen, T., Petkovic, M., Varga, O. E. & Jamuar, S. S. Editorial: Ai in healthcare: From data to intelligence. *Frontiers in Artificial Intelligence* **5** (2022). URL http://dx.doi.org/10.3389/frai.2022.909391.

[7] Onishchenko, D. *et al.* Screening for idiopathic pulmonary fibrosis using comorbidity signatures in electronic health records. *Nature Medicine* **28**, 2107–2116 (2022).

[8] Onishchenko, D. *et al.* Reduced false positives in autism screening via digital biomarkers inferred from deep comorbidity patterns. *Science advances* **7**, eabf0354 (2021).

[9] Onishchenko, D., Rubin, D. S., van Horne, J. R., Ward, R. P. & 65;6602;1c, I. C. Cardiac comorbidity risk score: Zero burden machine learning to improve prediction of postoperative major adverse cardiac events in hip and knee arthroplasty. *Journal of the American Heart Association* **11**, e023745 (2022). URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.121.023745.

[10] Sarda-Espinosa, A., Subbiah, S. & Bartz-Beielstein, T. Conditional inference trees for knowledge extraction from motor health condition data. *Engineering Applications of Artificial Intelligence* **62**, 26–37 (2017).

[11] Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)* **36**, 192–225 (1974).

[12] Gelfand, A. E. & Smith, A. F. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association* **85**, 398–409 (1990).

[13] Geman, S. & Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* 721–741 (1984).

[14] Centers for Disease Control and Prevention. Heart disease prevalence. https://www.cdc.gov/nchs/hus/topics/heart-disease-prevalence.htm. Accessed: [2024-01-15].

[15] 2022 alzheimer's disease facts and figures. *Alzheimer's & Dementia* **18**, 700–789 (2022). URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.12638. https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.12638.

[16] Imahori, Y. *et al.* Association of ischemic heart disease with long-term risk of cognitive decline and dementia: A cohort study. *Alzheimer's & Dementia* **19**, 5541–5549 (2023). URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1002/alz.13114. https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.13114.

[17] Omura, J. D., McGuire, L. C., Patel, R. *et al.* Modifiable risk factors for alzheimer disease and related dementias among adults aged greater than 45 years — united states, 2019. *MMWR Morb Mortal Wkly Rep* **71**, 680–685 (2022).

[18] Mendez, M. F. The relationship between anxiety and alzheimer's disease. *Journal of Alzheimer's disease reports* **5**, 171–177 (2021).

[19] Centers for Disease Control and Prevention. Data brief 405. https://www.cdc.gov/nchs/products/databriefs/db405.htm. Accessed: [insert date here].

[20] Xie, C., Wang, C. & Luo, H. Increased risk of osteoporosis in patients with cognitive impairment: a systematic review and meta-analysis. *BMC Geriatrics* **23**, 797 (2023).

[21] Song, Z. *et al.* Prevalence of Parkinson's Disease in Adults Aged 65 Years and Older in China: A Multicenter Population-Based Survey. *Neuroepidemiology* **56**, 50–58 (2021). URL https://doi.org/10.1159/000520726. https://karger.com/ned/article-pdf/56/1/50/3752546/000520726.pdf.

[22] Aarsland, D. & Kurz, M. W. The epidemiology of dementia associated with parkinson disease. *Journal of the Neurological Sciences* **289**, 18–22 (2010). URL https://www.sciencedirect.com/science/article/pii/S0022510X09008193. Mental Dysfunction in Parkinson's Disease.

[23] Liao, W.-C., Lin, C.-L., Chang, S.-N., Tu, C.-Y. & Kao, C.-H. The association between chronic obstructive pulmonary disease and dementia: a population-based retrospective cohort study. *European Journal of Neurology* **22**, 334–340 (2015). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ene.12573. https://onlinelibrary.wiley.com/doi/pdf/10.1111/ene.12573.

[24] Fragoso, C. A. V. Epidemiology of chronic obstructive pulmonary disease (copd) in aging populations. *COPD: Journal of Chronic Obstructive Pulmonary Disease* **13**, 125–129 (2016). URL https://doi.org/10.3109/15412555.2015.1077506. PMID: 26629987, https://doi.org/10.3109/15412555.2015.1077506.

[25] Bozkurt, B. *et al.* Heart failure epidemiology and outcomes statistics: A report of the heart failure society of america. *Journal of Cardiac Failure* **29**, 1412–1451 (2023). URL https://www.sciencedirect.com/science/article/pii/S1071916423002646.

[26] Li, J. M., Wu, Y. M., Zhang, D. M. & Nie, J. B. Associations between heart failure and risk of dementia: A prisma-compliant meta-analysis. *Medicine* **99**, e18492 (2020).

[27] Rajati, F., Rajati, M., Rasulehvandi, R. & Kazeminia, M. Prevalence of stroke in the elderly: A systematic review and meta-analysis. *Interdisciplinary Neurosurgery* **32**, 101746 (2023). URL https://www.sciencedirect.com/science/article/pii/S2214751923000294.

[28] Kuźma, E. *et al.* Stroke and dementia risk: A systematic review and meta-analysis. *Alzheimer's & Dementia* **14**, 1416–1426 (2018). URL https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2018.06.3061. https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2018.06.3061.