Winner's announcement

Information included in this section may be shared publicly with challenge results. If you are on a team, please complete the first two questions for each member of the team.

1. Please provide your preferred information for use in announcing the winners of the competition:
    a. **Name** (first and last name or first name and last initial): Prof Ishanu Chattopadhyay
    b. **Hometown**: Oak Lawn IL
    c. A recent picture of yourself or digital avatar (feel free to attach separately):
    d. **Social handle** or URL (optional): @ishanu_ch


2. Who are you (mini-bio) and what do you do professionally?

I am Prof. Ishanu Chattopadhyay, an Assistant Professor at the University of Kentucky's Institute of Biomedical Informatics in the Department of Medicine. I am an expert in biomedical informatics, general artificial intelligence, bleeding edge machine learning algorithms, and the computational aspects of data science. I lead innovative research at the intersection of AI and healthcare, developing predictive models and AI-driven tools to address complex medical challenges. My work has been recognized with prestigious awards and funded by various institutions, including DARPA, the NIH, and the Alzheimer's Association, and the research coming out of my group has been published in archival journals including Nature Medicine, PNAS, JAMA, JAHA, Science Advances. Persistent email: zeroknowledgediscovery@gmail.com

3. What motivated you to compete in this challenge?

The challenge presented a unique opportunity to apply my expertise in AI and machine learning to a critical healthcare problem. Alzheimer's Disease and Alzheimer's Disease Related Dementias (AD/ADRD) are areas where early diagnosis can significantly impact patient outcomes. The challenge of creating a generative AI framework to produce high-quality, accessible datasets aligned perfectly with my research goals of overcoming data accessibility barriers and promoting inclusive, data-driven healthcare solutions.

4. High level summary of your dataset: the data source, target, predictors, sample size, and use for early, inclusive prediction of AD/ADRD.

Our dataset, PhantomDB, comprises diagnostic histories of 2 million "phantom" patients generated using a novel generative AI framework. The data source includes time-stamped diagnostic and procedural codes for 21,374 patients from a national database and 187 African-American patients from the University of Chicago Medical Center, aged 60-75, who were eventually diagnosed with AD/ADRD. The target variables are diagnostic codes for AD/ADRD, and predictors include a wide range of ICD-10 diagnostic codes. This dataset facilitates early and inclusive prediction of AD/ADRD by providing a high-quality, representative dataset free from proprietary and privacy constraints.

5. What are two or three unique strengths of this dataset or type of data for early, inclusive prediction of AD/ADRD?

1. **Data Accessibility and Ethics:** PhantomDB circumvents data access barriers and aligns with ethical standards by replicating medical histories without infringing on proprietary data rights. This promotes a collaborative and inclusive research ecosystem.
2. **Demographic-Specific Modeling:** By focusing on both race-blind and African-American populations, the dataset addresses disparities in AD/ADRD incidence, advancing personalized healthcare and promoting equitable outcomes.
3. **Technological Innovation with "Phantom" Patients:** Our generative AI framework produces realistic medical histories, providing a high-fidelity replication of patient data free from proprietary constraints, thus enhancing the quality and applicability of the dataset for machine learning applications.

6. Did you use any tools or resources for developing your submission (e.g., to find a dataset, or explore the contents of a public dataset)?
We utilized proprietary datasets from MarketScan and the University of Chicago Medical Center for training our generative AI model. Additionally, we used various AI and machine learning tools, including Python and the **teomim** package available on PyPi, to develop and validate our phantom-net model. While we made use of proprietary data, it does not violate any data use agreements in place, since the data that is ultimately shared cannot be used to recreate in whole or in part the proprietary data.

7. Were there any data types or sources that you explored but didn’t fit for this challenge?
We explored integrating additional public health datasets and genomic data but found that they did not align perfectly with the challenge's focus on AD/ADRD prediction using clinical diagnostic histories. These data types were set aside to maintain the focus on developing a robust and representative dataset for early diagnosis.

8. How would you improve or enrich this dataset if you had access to a big research team and an unlimited budget?

With a larger research team and unlimited budget, we would:

A. **Integrate Multi-modal Data:** Incorporate genomic, lifestyle, and environmental data to enhance the predictive power and personalization of the dataset.

B. **Expand Demographic Coverage:** Include more diverse demographic groups, including different age ranges, socio-economic statuses, and geographic locations, to ensure the dataset's inclusivity and robustness.
C. **Longitudinal Data Collection:** Extend the timeline of patient histories to capture longer-term disease progression and more detailed comorbidity patterns, improving the dataset's utility for predictive modeling and early intervention strategies.