**Using large language models to enhance clinically-driven missing data recovery algorithms in electronic health records**

*Sarah Lotspeich, Wake Forest University*

Electronic health record (EHR) data are prone to missingness and errors. We previously devised an enriched chart review protocol using a "roadmap" of auxiliary diagnoses to recover missing values. However, chart reviews are expensive and time-intensive, limiting how many patients' data can be reviewed. We now investigate the accuracy and scalability of a roadmap-driven algorithm, based on ICD-10 codes, designed to mimic expert chart reviews and recover missing values. In addition to the clinicians' original roadmap, we consider new versions iteratively refined using large language models (LLMs) with clinical expertise to expand the list of auxiliary diagnoses. Using chart reviews for 100 patients from an extensive EHR, we examine algorithm performance. Across these patients, there were 413 missing values. Expert chart reviews recovered 49 (12%), while LLM-enhanced roadmaps recommended nearly twice as many (83-89, 20%-22%). The final algorithm, using clinician-approved LLM additions, offered a balance (73, 18%), expanding the roadmap with LLM suggestions only when clinically relevant. Applied to a larger study of 1000 patients from the same EHR, the per-patient median number of non-missing values increased from 6 to 7. Clinically-driven algorithms enhanced by LLMs can recover missing EHR data with accuracy comparable to chart reviews and can feasibly scale to large samples. Extending these methods to other data quality dimensions is a promising future direction.

**Translating Research from a Multi-Ethnic Asian Population Cohort**

*Marie Loh, Lee Kong Chian School of Medicine, Nanyang Technological University*

In this short talk, Assistant Professor Marie Loh will be sharing with us the opportunities and challenges that lies around the use of the rich phenotypic and molecular data available within the HELIOS-SG100K study, a multi-ethnic Asian population cohort, She will highlight the potential for translational opportunities, illustrating with several examples across different disease domains.

**Leveraging the All of Us Research Program for Longitudinal Disease Staging to Support Precision Medicine**

*Chih-Ting Yang, Department of Biostatistics, Vanderbilt University, Nashville, TN, USA*

Large-scale biobank resources provide new opportunities for precision medicine by integrating clinical, survey, genomic, and other biomedical data from diverse populations. In this talk, I will use the All of Us Research Program as an example to discuss how data science methods can help translate biobank data into longitudinal disease information.

As an example application, I will present a study of cardiovascular-kidney-metabolic syndrome, where we use electronic health record data to move beyond static baseline disease classification and characterize disease stages over time. This example highlights several key challenges in biobank-based research, including longitudinal staging, irregular clinical measurements, informative missingness, and interpretation of disease trajectories. More broadly, the talk will discuss how biobank data can support disease progression research, risk stratification, and precision medicine.

**Genetic Insights into Ocular Disorders and Comorbid Diseases: A Genome-Wide Association and Polygenic Analysis**

*Chiani Hsiung, Genetics Generation Advancement Corp.(GGA Corp.)*

Eye diseases, including cataracts, glaucoma, diabetes retinopathy, and age-related macular degeneration, are major global health challenges and leading causes of blindness. This study leveraged genome-wide association studies (GWASs) involving over 100,000 individuals, integrating data from the Taiwan Biobank and National Health Insurance Research Database, to identify genetic loci associated with disease onset. Our findings suggest that these conditions are influenced by multifactorial etiologies, as pleiotropic loci including rs10811660, rs4710941, rs2283228, and rs7646518 were identified, linking ocular diseases to metabolic conditions. Notably, a strong genetic correlation was observed between cataract and depression. Mendelian randomization analysis further demonstrated a causal effect of depression on cataract risk, implicating shared biological pathways, particularly oxytocin signaling, in disease pathophysiology. This finding revealed a functional genetic variant near the OXTR gene, highlighting its potential as a causal candidate for genetic diagnosis in precision health. By bridging the gap between genetic discovery and clinical application, this research offers critical insights into shared genetic mechanisms across diverse health domains, paving the way for innovative diagnostic and therapeutic strategies.