

International Journal of Population Data Science

Journal Website: www.ijpds.org



Implementing a Scalable, Secure Genomics Ingest and Processing Pipeline in a Trusted Research Environment for Dementia Research

Alieyeh Saraband Moghaddam¹, Lewis Hotchkiss¹, Emma Squires¹, and Simon Thompson¹

¹Dementias Platform UK, Swansea, United Kingdom; Swansea University, Swansea, United Kingdom

The integration of genomics data into dementia research offers transformative potential for understanding disease mechanisms and enabling precision medicine. However, its utility is constrained by significant challenges in data sharing, privacy, standardization, and computational scale. To address these barriers, we designed and implemented a secure, novel and scalable genomics ingest and processing pipeline within the Dementias Platform UK (DPUK) Trusted Research Environment. Our comprehensive, end-to-end framework encompasses a seven-stage process, including cohort discovery via an interactive metadata matrix, rigorous quality and privacy controls using bioinformatic tools (e.g. PLINK, VCFtools, etc.), a genomic data organization standard for scalability and balance between interoperability and flexibility, as well as secure provisioning within a Five Safes governance model. A key innovation is the use of a MinIO-based object storage architecture with custom metadata tagging, enabling efficient, privacy-

preserving data access at large scale. We also operationalized a standardized polygenic risk score (PRS) pipeline to generate validated, derived datasets. This integrated system now facilitates secure research access to harmonized genomic data across 15 diverse cohorts, including population-based, clinical, and family studies, significantly expanding the resources available for dementia research. This work presents a replicable and governance-first model for the responsible management of complex, high-volume linked data. Looking ahead, we plan to migrate key computational modules, particularly the PRS generation and quality control workflows, to Nextflow. This migration will enhance computational reproducibility, enable portable and parallelized execution across high-performance environments, and further standardize analytical processes, directly advancing the methodological innovation and scalability goals of population data science.

