

International Journal of Population Data Science

Journal Website: www.ijpds.org



Developing Synthetic Data Tools for Trusted Research Environments to Enable Researcher Training

Lewis Hotchkiss^{1,2}, Kafayat Adeoye^{1,2}, Emma Squires^{1,2}, and Simon Thompson^{1,2}

¹Dementias Platform UK Data Portal, Swansea, United Kingdom

²Swansea University, Swansea, United Kingdom

We aimed to develop an open-source Python package for the Trusted Research Environment (TRE) community to facilitate the generation of varying levels of synthetic data. One of the main use cases of this tool is to support the generation of synthetic data which can be used for researcher training.

Our tools enable users to generate synthetic datasets with low, medium or high fidelity based on real-world data. This is achieved through using both statistical techniques and machine learning to generate synthetic data with varying degrees of quality and privacy guarantees. It includes built-in evaluations for measuring privacy, quality and utility, as well as automated report creation, ensuring transparency to enable data owners to make informed decisions for releasing synthetic datasets from TREs. This addresses the challenge of researcher training in TREs, where strict governance controls often hinder access to real data.

This tool has successfully been implemented by the Dementias Platform UK (DPUK) Data Portal, with further plans to deploy across other TREs within SeRP. We have generated several synthetic versions of DPUK datasets - both lower fidelity with strong privacy guarantees and higher fidelity which retains essential statistical properties. The evaluation framework ensures that the synthetic data we generate meets privacy and utility thresholds, making it suitable for researcher training and method development. Additionally, we have been working on the governance to be able to support the deployment of these synthetic datasets into practice. Overall, our synthetic data significantly enhances training opportunities, overcoming governance restrictions on real data.

Our open-source Python package addresses the challenge of researcher training in TREs by generating synthetic data with built-in privacy, quality, and utility assessments. This enables secure and practical training without compromising governance controls. By facilitating synthetic data generation, evaluation, and reporting, our tool enhances research capacity.

