

New synthetic datasets to assist COVID-19 and cardiovascular research

The Medicines and Healthcare products Agency (MHRA) has announced the creation of [two innovative synthetic datasets](#) which will support the development of cutting-edge medical technologies to fight COVID-19 and cardiovascular disease.

These datasets have been generated to accurately mirror symptoms, diagnoses and treatments in genuine patients. They are based on anonymised primary care data using innovative methods to produce entirely artificial data that doesn't contain any original data from 'real' patients, further reducing risks to patient privacy.

Synthetic datasets like these are valuable in the development and testing of machine learning and artificial intelligence (AI) algorithms in medical devices used for diagnosing diseases and monitoring and improving health conditions.

CPRD Director Janet Valentine comments:

These datasets are designed to help researchers and companies validate their innovative new AI and medical devices. This development will support bringing safe products to market sooner, enabling patients to benefit from the latest technical advances.

The datasets were produced by a collaboration between the Clinical Practice Research Datalink (CPRD), MHRA Medical Devices Division and researchers at Brunel University.

The synthetic data generation methodology and the cardiovascular dataset were funded by a grant from the [Regulators' Pioneer Fund](#) launched by The Department for Business, Energy and Industrial Strategy (BEIS) and managed by Innovate UK. Creation of the COVID-19 synthetic dataset was funded by NHSX.

Indra Joshi, Director of AI at NHSX, said:

At NHSX we are committed to protecting patient privacy whilst supporting the development of cutting-edge technologies that could potentially help the NHS and our patients. "Creating synthetic datasets is a novel way to help train machine learning algorithms on a rich and diverse set of data whilst maintaining safety and protecting privacy.

The data generation and evaluation framework, as well as the datasets, are owned by the MHRA. A detailed technical description of the methodology used

to generate the synthetic datasets is available [here](#).

For access to these datasets, please submit [an application form](#) to enquiries@cprd.com including 'Synthetic data access request' in the email subject header. Applicants from organisations that are not existing CPRD clients will also need to submit [a new client request form](#).