PROCESS – DPUK DATA MANAGEMENT

Background

This document provides details of the procedures that underpin the DPUK Data Management Policy, found on the DPUK website.

Procedures

Governance

Cohorts are invited to deposit data with DPUK. This can be as much, or as little, of their data with complete and partial data transfers supported.

Transfer of data to DPUK does not imply any change in legal ownership or governance principles of any dataset. DPUK acts as an enabling infrastructure, working with data owners to facilitate the curation and storage of data, and access to data. DPUK invites, but does not require, cohorts to use its central data access application system.

Deposited data may be removed at any time, without providing cause, by receipt by UKSeRP acting on behalf of DPUK, of a mailed letter with wet signature from the data owner requesting the removal of data from DPUK. Emailed correspondence of such a request shall not be binding, although notice of such written request can be supplied in such a format. Cohort data will be removed from UKSeRP by Swansea once all studies using the data have completed, as set out in the Data Deposit Agreement. If no studies are ongoing, data removal can be immediate.

Upon transfer or at an appropriate stage of discussions surrounding cohort data, a refresh rate can be established in order that data are kept up to date as necessary in line with ongoing cohort work.

Primary Data Transfer

Cohort data will be transferred to the DPUK instance of UKSeRP, hosted by Swansea University, to be prepared for access by the DPUK research community. Data preparation will be by permitted persons nominated by the cohort upon signature of the DPUK Data Deposit Agreement.

DPUK will receive cohort data ‘as is’ (native), and will curate the data according to a DPUK common data model. Data curation will include liaison with the cohort research team to ensure the sense of the data is retained. Researchers may request datasets in native or curated format.

Once datasets are migrated to the UKSeRP, they will be mapped enabling the relevance of datasets for specific research questions to be ascertained. Bespoke cohort datasets are created according to application form criteria selection. Datasets will be mapped using variable and participant selection tools, which assume a two-dimensional data matrix of columns (variables) and rows (participants). High dimensional data will be subject to pre-processing by specialist tools. The variable selection tool enables the availability of individual variables within each dataset to be identified. The participant selection tool enables the number of participants with data on each variable and with data on combinations of variables to be identified: for example, the number of participants in a cohort within a certain age bracket and with a family history of dementia.

Upon transfer, a dataset will enter into a split-file anonymisation process, to reduce the risk of identifying individuals. This process is undertaken with NHS Wales Informatics Service (NWIS), a Trusted Third Party (TTP) of Swansea University, who are contracted to perform one half of the split-file anonymisation.

The process is briefly detailed as follows:

- File 1 includes all identifiers held in the rowed data, for example NHS number, database unique identifier and date of birth. This file is sent (by the cohort or Swansea depending on agreed method of transfer) to NWIS, where it is encrypted. All identifiers are at this stage changed to a separate unique

identifier called an Anonymous Linking Field (ALF). Date of birth reverts to week of birth for extra separation. This newly encrypted file is then sent to Swansea by NWIS.

- File 2 includes all clinical and administrative information present in the particular study. This will contain all variables that are not present in File 1, and for the purposes of DPUK, all data can be transferred. This file is sent directly to Swansea, where it will also be encrypted. In order to match File 1 and File 2 back together to create the fully anonymised data rows, File 2 will also need to contain a unique identifier, which will be as suggested above, a uniquely created system ID.

- Swansea match both File 1 and File 2 back together once encrypted in order to create a 'File 3', known as an ALF-E. This file will contain the exact data provided by the cohort; however it will now be in standardised format for the analysis platform having been anonymised via double encryption.

If data are anonymised at source before transfer, DPUK can be given access to the data to facilitate the encryption process via a method to be discussed on a cohort to cohort basis, most usually by DPUK providing access to nominated cohort staff to the Research Data Appliance, a front-end data management tool for linking data to the infrastructure. Other methods of transfer depending on cohort preference could include, for example, the creation or granting of access to an FTP (File Transfer Protocol) server; the process is flexible according to the needs of the cohort.

Data Storage

Cohort data will be housed in its own dedicated area of the UKSeRP, readily accessible by cohort permitted persons, and upon request and subject to access procedures by the wider DPUK research community. This dedicated area will be managed via the Research Data Appliance from a data upload and permissions point of view and raw data accessible on a Virtual Desktop Infrastructure (VDI) via two-factor authentication (username and password plus mobile device authentication). The VDI houses software such as SPSS and R, although the space can be customised for any one user's/study's requirements (subject to additional costs).

Data held by DPUK will only be made available to bona fide academic and industry researchers who have successfully completed the DPUK study application process. Bona fide researchers are defined as being any one researcher with professional expertise to conduct bona fide research, and who has a formal affiliation with a bona fide research organisation that requires compliance with appropriate research governance and management system

The web-facing part of the DPUK Data Portal is automatically accessible to those users registering with an academic email address, accessible to industry partners once their company and email has been verified, and is not accessible to users with a universal public email address (eg Gmail, Hotmail etc.).

Data Use

Data provided to researchers will be available within a VDI, rather than a web-portal interface. In this situation, data will always be geographically located in DPUK servers (in Swansea) from anywhere in the world in which the data are being accessed. This explicitly restricts data manipulation within the virtual desktop. Data will be provided within a file folder, which will be shared amongst nominated users should the project involve multiple approved users.

Analysis of data will be performed within the VDI, using the supplied analysis and statistical software, which can be customised (subject to additional costs).

Originally supplied individual row-level data and subsets of such data are not permitted to leave the VDI. Results are restricted to certain file types such as SQL scripts, SPSS syntax, SPSS outputs, frequency counts, statistical outputs and project reports. Such results are scrutinised before approval to be released from the VDI. (Please note that this approvals process is subject to change, where certain files are automatically allowed out of the VDI and the process regularly audited.)

Results subsequently used in publications will be reviewed as part of the publication process and policy. Access to the DPUK Data Portal Analysis Platform will be governed by the DPUK Data Access Procedure.

Release date of document: 26 October 2020