Working with big data

Dementias Platform UK
Medical Research Council

Aims and objectives

The aim of this user guide is to provide information and guidance for researchers who are working with large datasets for the first time.

Specifically, this user guide will:

- provide general advice for working with large datasets;
- provide advice and guidance for data preparation and cleaning;
- suggest general data processing and analysis guidance.

General

- Working with large datasets is no more complicated than working with smaller datasets
- It still requires a strict systematic approach
- It often requires learning a new statistical software package: SPSS and Excel do not have the capacity or functions for big data analytics
- Invest in courses or watching online tutorials and videos (eg Stata, SAS, R, Python)
- Invest time in understanding longitudinal methodologies (if relevant), eg replenishment of participants, withdrawals, statistical methodologies

Data preparation

- Collaborate: the Data Portal provides the opportunity to share your workspace with other members of your team
- Use all resources accompanying datasets, eg data dictionaries, cohort profile papers, existing journal articles, code books
- Invest time in reading and researching original questionnaires where relevant and available
- Maintain electronic notetaking throughout for audit and publication purposes

Data cleaning

- Work using scripts and electronic notebooks – do not attempt to change the master dataset or duplicate a copy
- Investigate using derived variables and existing algorithms
- Understand the variables and the differences between each individual variable (eg ‘slices of bread eaten’ vs ‘slices of bread eaten with spread’)
- Time spent cleaning the data will be time well spent when it comes to processing the data
- Carefully consider the missing data and how they are coded – determine how they will be replaced

Data processing and analysis

- Use a systematic labelling system and clearly defined scripts which outline processes and procedures (eg a ‘Do-file’ in Stata)

- Consider big data methods such as structural equation models, machine and deep learning, and psychometric methodologies

- Consider individual cohort coding differences (eg how is gender coded across the individual cohorts?)

- Research the interpretation of output in large datasets (ie look beyond the *p* values which may all be significant due to sample size)

- Consider testing scripts and code on smaller subsets first

Dementias Platform UK
Medical Research Council

There is more support available for you online

www.dementiasplatform.uk/supportforresearchers

If you have any questions, please get in touch

dpuk@psych.ox.ac.uk | @DementiasUK