Challenges of working with a large database of routinely collected health data: Combining *SQL* and *R*

Joanne Demmler[1], Caroline Brooks[1], Sarah Rodgers[1], Frank Dunstan[2], Ronan Lyons[1]

1. Swansea University, College of Medicine, Grove Building, Singleton Park, Swansea SA2 8PP
2. Cardiff University, Department of Primary Care & Public Health, Neuadd Meirionnydd, Heath Park, Cardiff CF14 4YS
*Contact author: j.demmler@swansea.ac.uk

Keywords: SAIL, anonymised data, RODBC, child health, data linkage

Introduction: Vast amounts of data are collected about patients and service users in the course of health and social care service delivery. Electronic data systems for patient records have the potential to revolutionise service delivery and research. But in order to achieve this, it is essential that the ability to link the data at the individual record level be retained whilst adhering to the principles of information governance. One such example is the Secure Anonymised Information Linkage (SAIL) databank, which contains health, social and education data for three million residents for a contiguous area in Wales, UK. There are currently 21 major datasets containing about 1.6 billion records, which can be linked anonymously at the individual level and household level.

Background: All work on SAIL data is executed through a secure remote desktop environment via a virtual private network (VPN), which has no internet connection and all output requires approval before it can be released for external viewing or publication. The processor speed equals 1 core of a Xeon X5550 @ 2.67 GHz processor, with an allocated memory of 2GB RAM per user.

Methods: We present here an example from the National Community Child Health Database, which contains height and weight measurements from school entry examinations for 849,238 children. Data are preselected using *SQL* and saved as a temporary table in SAIL to remove children with negative age at examination and to restrict examinations to the years 1990 to 2008. After removal of biologically unfeasible records with *SQL* 1,764,728 records for 594,720 children are imported into *R* using the `sqlfetch` command of the **RODBC** package. A simple algorithm is explored to remove remaining outliers in the data.

Results: Although *R* is very effective in some basic analysis and in the exploration of the data, it is not very efficient in dealing with such a vast dataset in certain situations. Memory limitations within the secure gateway platform mean that quite simple *R* scripts might run for a considerable time (days) and data might get lost in transfer operations (saving back to SAIL might fail depending on the table dimensions). At the present, this prevents the usage of more advanced statistical methods as well as modelling of the data.

Outlook: As a result of this analysis we have decided to migrate *R* onto a more powerful server. We are also investigating the possibility of multithreading *R* code, using the College of Medicine's BlueC replacement supercomputer (2 nodes, each with 30 Power7 cores @ 3.3GHz and 100GB of Ram).