Safe Data Access Professionals Topic Event
Statistical Disclosure Control (SDC) automated tools

Tuesday 7th September 2021, 10.00 – 12.30
Virtual meeting

Attendees

Christine Woods (UK Data Service) (Chair), **James Scott** (UK Data Service) (Deputy Chair), **Helen Cadwallader** (UK Data Service) (SDAP Membership and Administration Officer), **Ramiro Bravo** (University of Manchester), **Olly Butters** (University of Liverpool), **Mary Cleaton** (Office for National Statistics), **James Dainty** (HM Revenue & Customs), **Jorgen Engmann** (The Health Foundation), **Jara Kampmann** (GESIS), **Beate Lichtwardt** (Social Sciences and Humanities Open Cloud), **Elaine Mackey** (University of Manchester), **Simon Parker** (Deutsches Krebsforschungszentrum), **James Rayner** (UK Data Service), **Beth Routley** (Office for National Statistics), **Aida Sanchez** (Centre for Longitudinal Studies), **Tony Stone** (University of Sheffield), **Alix Taylor-Scopes** (UK Data Service), **Amy Tilbrook** (University of Edinburgh), **Soyful Ullah** (HM Revenue & Customs), **Anca Vlad** (Cancer Research UK).

Apologies

Emily Griffiths (University of Manchester).

Minutes

1. Welcome and Introductions

The chair welcomed all to the event, including new members (6). Everyone then introduced themselves.

2. Statistical Disclosure Control (SDC) automated tools

Olly Butters (University of Liverpool) delivered a presentation on 'An overview of DataSHIELD'.

From the DataSHIELD website: "…DataSHIELD provides a novel technological solution that can circumvent some of the most basic challenges in facilitating the access of researchers and other health care professionals to individual level data. Although initially developed for work in the biomedical and social sciences, DataSHIELD can be used in any setting where microdata (data on individual subjects) must be analysed but cannot physically be shared with the research users.

DataSHIELD is a flexible, modular, free, open-source solution ideally placed to grow a broad user and development community."

Presentation slides to accompany these minutes is available here. A recording of the presentation is available here.

Following the presentation there were questions and discussion.

- OB confirmed DataSHIELD was used in many cohort studies, up to 20, mostly in Europe, including a large-scale European cohort, EUCAN-CONNET, and, in Canada, where the automated tool was applied in analysis across country boundaries. DataSHIELD is also being used by a UK wide consortium, LifeCycle, across several sites. For example, a group in Germany are analysing hospital data by creating a wrapper that sits outside the firewall where data requests are held, whilst inside the firewall queries are raised on any new requests held in the extension and pulls these in. Whilst in Spain, there is a genomic project, Bioconducter, that is using DataSHIELD.

- The group learned how the broader issue of secondary disclosure control is currently handled in DataSHIELD as a historical record of every research command on all sets and subsets of data ever requested. A range of real time solutions are being considered and this may be an enquiry for future consultation with the SDAP network community.
- The group also recognised the future potential of DataSHIELD as automated tool in supporting the process of cross checking outputs, with reference to the original research request and also against outputs generated by other researchers using the same data. This is one of the future development goals for this tool.
- The group also reflected on the matter of vertical partitioning and data linkage. OB confirmed that a governance solution was applied, where data is used as it has been provided and that vertical partitioning and data linkage was not often used as a solution.
- DataSHIELD has been written using open source, and therefore, free, although a staffing resource is required to install and maintain it, a function that could be adopted within an existing team. The tool has been written in 'R'. Currently, if data is requested from a source with tabulated data, should results fall below a given threshold, the data would be withheld and a new command would need to be raised.
- Future developments also include installing DataSHIELD within a TRE and building in commands with permissible outputs. To date, there is no automated solution to the underlying matter of data harmonisation and the group concurred that standardisation of meta data is major, ongoing issue.

Useful links provided during the session + Q&A

- **DataSHIELD:** secure bioscience collaboration - https://www.datashield.org/events for the upcoming DataSHIELD Conference (online) 10-11 November, 2021.
- **EUCAN-Connect** (173 cohorts, EU funded project): - https://eucanconnect.com
- **RECAP** (formerly **LifeCycle** and with a significant UK component, EU funded project): 30 year longitudinal study focused on preterm births with the aim to improve the health, development and quality of life of these children and adults by optimizing the use of population data in policy development - https://recap-preterm.eu/about-recap-preterm/members/

Applied in the following projects as mentioned in Q&A:

- **MIRACUM:** medical informatics in research and care in university medicine (Germany based, EU funded project) – https://www.miracum.org/
- **Bioconductor:** open source software for bioinformatics (Spain) - https://www.bioconductor.org
- **Full list of projects** detailed here, scroll down - https://www.datashield.org/about

3. SDAP members' experiences of using SDC automated tools

There was a brief discussion about SDAP members' experiences of using SDC automated tools.

4. Service Updates

Service updates were given by:

- **James Dainty** (HM Revenue & Customs)

5. AOB and close

The chair thanked the speaker for his presentation and all attendees for their contributions to the meeting. The chair then closed the meeting.

The next SDAP Topic Event ‘Review and update of SDAP Competency Framework’ will be on **Tuesday 7th December, 2021**.