International Journal of Population Data Science

Journal Website: www.ijpds.org

Can censoring of research outputs be automated to ensure robust data protection?

Nicholas, Michael¹, Davies, Chris¹, Nock, Kelly^{1*}, Bailey, Kerry¹, Barker, Craig², Player, Luke³, and Thomas, Helen²

¹We Predict

²Abertawe Bro Morgannwg University Health Board

³Swansea University

Background

Guidance regarding research outputs recommends censoring so that even when aggregating anonymised linked data no cell should have less than 5 -10 units. This is recommended to decrease the likelihood of re-identification. Leaving those cells empty is not adequate if other cells can be used to identify the numerical value of that cell. Some outputs necessitate a large number of tables to be exported this will become more common. This was the case where the outputs from a research study involved several large tables which drove a front end interactive visualisation. As linked data outputs are used to make operational decisions which necessitates timely data outputs of large amount of aggregated data this issue will be more common. Human scanning of all tables may not be time or cost effective and can be subject to human error.

examples of the impact on the output. Some stakeholders felt that the censoring of the anonymised aggregated data went beyond the 'reasonable effort' required to re-identify individuals. Some expressed the opinion that the lack of detail and missing data that this method results in is excessive and has been sacrificed for the sake of minimal risk. Some stakeholders felt the risks had been allowed to outweigh the societal benefits. The team were assured that although the censoring may be considered excessive by some it did ensure safe censoring and offered as low a risk a possible for re-identification. However routine implementation of this method has not been agreed.

Approach

Many methods of censoring were considered including Barnardisation (adding or subtracting 1 randomly to small numbers) suppression and a combination of methods. It was then necessary to code the methods to ensure that censoring was implemented in all cells in the output and that the output was still meaningful. It was then necessary to check the outputs for quality and introduce an 'audit' system to ensure that the quality was maintained but did not impact on the outputs of the findings.

Discussion

Software engineers were able to develop an algorithm that performed safe censoring using a level of '10 or under'. It also ensured that the statistical tables were still functional. The presentation will describe how this was done and demonstrate some

*Corresponding Author:

Email Address: knock@wepredict.co.uk (K. Nock)