International Journal of Population Data Science

Journal Website: www.ijpds.org

Developing data governance standards for using free-text data in research (TexGov)

Jones, K[1*], Ford, E[2], Lea, N[3], Griffiths, L[1], Heys, S[1], and Squires, E[1]

[1] Swansea University

[2] Brighton and Sussex Medical School

[3] University College London

Background

Free-text data represent a vast, untapped source of rich information to guide research and public service delivery. Free-text data contain a wealth of additional detail that, if more accessible, would clarify and supplement information coded in structured data fields. Personal data usually need to be de-identified or anonymised before they can be used for purposes such as audit and research, but there are major challenges in finding effective methods to de-identify free-text that do not damage data utility as a by-product. The main aim of the TexGov project is to work towards data governance standards to enable free-text data to be used safely for public benefit.

Methods

We conducted: a rapid literature review to explore the data governance models used in working with free-text data, plus case studies of systems making de-identified free-text data available for research; we engaged with text mining researchers and the general public to explore barriers and solutions in working with free-text; and we outlined (UK) data protection legislation and regulations for context.

Results

We reviewed 50 articles and the models of 4 systems providing access to de-identified free-text. The main emerging themes were: i) patient involvement at identifiable and de-identified data stages; ii) questions of consent and notification for the reuse of free-text data; iii) working with identifiable data for Natural Language Processing algorithm development; and iv) de-identification methods and thresholds of reliability.

Conclusions

We have proposed a set of recommendations, including: ensuring public transparency in data flows and uses; adhering to the principles of minimal data extraction; treating de-identified blacklisted free-text as potentially identifiable with use limited to accredited data safe-havens; and, the need to commit to a culture of continuous improvement to understand the relationships between accuracy of de-identification and re-identification risk, so this can be communicated to all stakeholders.

*Corresponding Author:
Email Address: k.h.jones@swansea.ac.uk (K Jones)

https://doi.org/10.23889/ijpds.v4i3.1332