

International Journal of Population Data Science

Journal Website: www.ijpds.org



Automating Low-Fidelity Synthetic Data Generation from Metadata: A Reproducible Approach Using R

Linus Chirchir¹

¹Research Data Scotland, Edinburgh, United Kingdom

Objectives

Access to pseudonymised administrative data for research is very important for privacy preservation and regulatory compliance. Synthetic data (SD) offers a solution to address concerns related to accessibility, security and privacy. We therefore developed a scalable framework that automates the generation of low-fidelity synthetic data (LFSD) from publicly available metadata.

Methods

Our R-based automated and reproducible pipeline takes metadata as input and generates SD that does not maintain the relationship between the observed variables in[SM1] the original dataset. The pipeline imports metadata from datasets (data dictionaries and lookup tables), generates pre-defined pattern-based categorical, numeric and date-time variables and applies missingness on pre-defined specifications. It also randomises the SD generated, preventing disclosure; an associated quality report performs validation checks to ensure that[SM2] the structure of SD is as similar as possible to the real data (RD) whilst minimising disclosure risk. The tool has been used in successfully generating four education synthetic datasets. [SM3]

Results

The framework successfully generated LFSD that closely mimicked RD while protecting individuals' sensitive information. To avoid any real or apparent data breaches, SD will be checked based on four parameters: labelling (to ensure SD was clearly identified as synthetic and not RD), disclosure (disclosure risk evaluation (DRE) carried out and any risks identified were mitigated), structure (ensured structure of SD is as similar as possible to RD) and documentation (differences in the structure of SD compared to RD data were documented). We found that SD maintained the cardinality of categories (within $\pm 5\%$) with RD. There was no explicit identity exposure as validated by the generation of a dynamic SD quality assessment report which is shared with the researchers.

Conclusion

This SD pipeline ensures protected access to administrative datasets while upholding confidentiality. The automation of the framework streamlines workflow processes, minimises manual effort and enables timely and cost-effective access to SD by researchers. Future work could look at refining fidelity options and developing an R package for broader usability.

