

International Journal of Population Data Science

Journal Website: www.ijpds.org



What are the useful characteristics of synthetic data? A qualitative study of researcher perspectives

Elizabeth Pattinson¹ and Emily Oliver¹

¹ADR UK, London, United Kingdom

Researchers may consider using synthetic data in place of real data when they face limited data access and lengthy approval procedures to conduct research with the data. In order to optimise the practice of creating and using synthetic data, it is important to understand what characteristics will be of most use. We found limited insight on this in existing literature and conducted our own qualitative study to explore it further. The objective of our study was to learn from researchers which aspects of synthetic data are important for utility, in what contexts they see benefits and what is needed to enable optimal use of synthetic data going forward. Informed by real data quality frameworks, we structured interview questions across different aspects and recruited 19 researchers from a variety of sectors, disciplines and career levels for interview. We used

a semi-structured interview protocol and thematic analysis of interview transcripts to uncover four areas of interest: barriers to real data access, versatility of synthetic data use cases, dependencies that affect synthetic data adoption, and the underdeveloped nature of the ecosystem. Our research revealed useful synthetic data characteristics identified by researchers covering accessibility, quality and safety. Balanced against these are the need for proportionate governance and cost, as well as clear and explicit guidance on responsible use. This research enabled us to develop the outline of a synthetic data characterisation framework that provides insights about user expectations alongside recommendations for synthetic data creation and provision.

