

International Journal of Population Data Science

Journal Website: www.ijpds.org



Guiding principles for the provision of synthetic data

Emily Oliver¹

¹ADR UK, London, United Kingdom

Synthetic data is an emerging tool for those learning to use and/or planning to use secure datasets, and for collaborators on secure data projects. However, there is a lack of consistency in its provision, and this affects both trust in its use and appetite to routinely produce it as a resource for researchers. There are important considerations for data owners and providers to ensure they provide synthetic data that meets the needs of users, remains compliant, and has overall positive impact. As the synthetic data field is still relatively new, norms and precedents are yet to fully emerge and develop, and standards and consistent approaches are lacking. We collaborated with experts across a range of sectors and drew on recent research with stakeholders to unpack the key facets of synthetic data provision, including worries and realities. Starting with a

focus on synthetic data with low disclosure risk (typically low fidelity), we created a structured workflow that involves planning, creating, validating and releasing a synthetic dataset. We then used it to develop guiding principles for providers framed around a 'Plan, Do, Check, Act' model. Each stage includes specific steps to ensure clarity, consistency, quality and effective risk control. By sharing the principles and the steps to support implementation, we anticipate that this approach can be helpful in informing organisational policies on synthetic data provision. We also hope that further discussion, debate and testing will lead to increased engagement, trust and a more consistent approach to creating and sharing synthetic data.

