Synthetic data can be thought of as artificial data that closely mimic the properties and relationships of real data. The concept is not new but recent technological advances, including machine learning methods, have seen a rapid growth in interest in synthetic data including potential applications in health and life sciences. Some use cases are based on reducing the privacy risks and burden of legal compliance that would be in place for sharing and processing real personal or private data. However, synthetic data can also be generated to replace missing information needed to test products, software or sections of code, and to train or validate AI tools. It is this potential which has led the Medicines and Healthcare products Regulatory Agency (MHRA) via its Clinical Practice Research Datalink (CPRD) service, to develop expertise in the creation of synthetic datasets, resulting in a number of synthetic datasets that can be used for training purposes or to improve algorithms or machine learning workflows.

The pace of technical progress is outstripping regulatory guidance and it is unclear whether, or under what conditions, synthetic health data will be considered ‘personal data’ governed by data protection law (the UK GDPR and EU GDPR). In this report we seek to respond to this uncertainty and identify whether, or in what circumstances synthetic health data are ‘personal data’ through consideration of technical approaches to synthetic data generation and analysis of relevant law, guidance and academic commentary in the UK and EU.

It is important to recognise that there are a wide range of synthetic data methods and technologies that can be used to generate different forms of output data, from manual generation based on expert knowledge, to iterative manipulation of real data, through to fully automated generation using machine learning methods like generative adversarial networks. Output datasets may also be partially or fully synthetic, and generated for a wide range of purposes. This means that there is no one-size-fits-all answer to the question of whether synthetic data are ‘personal data’ and it will be the responsibility of each data controller, in consultation with developers and users, to evaluate the legal status of input and output datasets in context.

Our legal analysis highlights that regulators and the courts are yet to grapple fully with synthetic data generation and that data authorities across the EU and UK are cautiously positive about the potential of synthetic data while recognising evidence of potential privacy risks. However, we can discern what could be termed an ‘orthodox’ approach to synthetic data being adopted by the regulators. This views synthetic data as a novel privacy enhancing technology (PET) and begins with the position that if the input or training data are ‘personal data’ it is presumed that models and output data will remain personal data unless effective anonymisation can be demonstrated with confidence.
Unfortunately, demonstrating that data are no longer ‘personal data’ is not straightforward. Any assessment needs to be comprehensive and involves careful scrutiny of a wide range of factors including the nature of the data itself, the range of risks and attacks that could threaten the synthetic dataset or a synthetic data model and the technical and organisational safeguards in place to protect the data environment. Although such an assessment is inherently subjective and there is no single defined threshold of identifiability which can be used, best practice incorporates using quantifiable statistical assessments where feasible as well as conducting penetration or motivated intruder testing. An audit or data protection impact assessment can assist in identifying appropriate additional safeguards which may be required. Ultimately, an assessment may be made that the risk of identification is so low (remote or negligible) that the synthetic data or model do not constitute personal data. However, a controller is obliged to keep this under review and adjust the assessment and apply further safeguards if new threats, technologies or additional sources of information arise which could increase the risk.

While this ‘orthodox’ approach may ensure privacy risks are minimised, there are challenges that regulators and policymakers should bear in mind. First, it is likely to lead to risk-averse conclusions and decisions to, for example, reduce the utility of data through use of techniques like differential privacy or, tightly limit access to synthetic data and restrict how it may be used. Second, this places a high burden on synthetic data developers (and potentially users, who may also have to make an assessment of privacy risks) due to the time and expertise required to fully audit identification risks and make consequent adjustments to the data or the data environment. This could slow the availability of synthetic data for health research and development purposes or suppress access to synthetic data for health and research purposes if costs are passed on to users.

Third, if it is genuinely determined that synthetic datasets and models constitute ‘personal data’ this gives rise to significant complexity in determining how data protection rights and obligations might apply. For example, how does the principle of data accuracy and right to rectification contained in Article 16 UK GDPR apply to synthetic data where it is not even clear that a relevant individual is the focus of the data? To whom must information be provided under Articles 13-14 UK GDPR? How does a right to object to processing apply if an individual’s data has at some point been used to develop a model? Can a model ‘unlearn’ that information?

There may also be reasons to consider adopting a different approach to some forms of synthetic health data generation. This could be because the nature of some data synthesis techniques and models in practice results in almost negligible identification risks and/or it is inappropriate to view remaining risks (such as coincidental creation of synthetic data that match a real human who was not even part of the training data)
Executive summary

as central to data protection law. This touches on a wider debate about whether the concept of ‘personal data’ has been overstretched, giving rise to ever greater complexity and uncertainty for data subjects and controllers. Some scholars argue that a more appropriate approach to would be to limit the threshold for ‘personal data’ but bring forward regulation of aspects of algorithmic processing to address the range of potential harms, including many that are not addressed by data protection law, such as group harms.

Now would be good time for regulators (in particular the Information Commissioner’s Office), health data authorities, technical experts and legal specialists to consider the regulatory approach to synthetic health data generation and whether a different model, such as a shifted presumption that some forms of synthetic data are non-personal data, may be more proportionate and technically feasible in certain circumstances.

Recommendations

Throughout this report we highlight specific considerations for synthetic data developers, researchers, regulators or policymakers. We also make three overall recommendations:

1. synthetic data developers and users should continue to follow best practice in relation to data protection impact assessments and anonymisation in assessing the identifiability and other data protection risks arising from processing.

2. synthetic data developers, researchers, regulators and policymakers should seek to achieve greater clarity, and reach consensus on:
   a. appropriate standards and approaches to assessing identifiability of specific synthetic data generation methods, utilising quantitative metrics as far as possible;
   b. whether the default for regulating certain forms of synthetic data and synthetic data generation should change from presumptively ‘personal data’ to a more proportionate approach that allows for some synthetic data to be classified as non-personal data based on an assessment of risk by data controllers.

3. as synthetic data generation and other forms of AI-driven processing for health purposes gain pace, regulators and policymakers should prioritise determining what form of regulation is appropriate for this sector and how it fits within the overall regulatory framework.