Are synthetic health data ‘personal data’?

PHG FOUNDATION
making science work for health

Authors

Colin Mitchell & Elizabeth Redrup Hill

May 2023

Are synthetic health data 'personal data'? A PHG Foundation report independently commissioned by the MHRA to assess the status of synthetic health data in UK data protection law.

Acknowledgements

We are grateful to Puja Myles & John Latham-Mollart from the CPRD and Richard Branson from the MHRA for their valuable input and assistance in this research.

Disclaimer

The following report is intended to provide general information and understanding of the ethical and legal framework. It should not be considered legal advice, nor used as a substitute for seeking qualified legal advice.

URLs in this report were correct as of March 2023

Written and produced by PHG Foundation

2 Worts Causeway, Cambridge, CB1 8RN, UK +44 (0)1223 761900

www.phgfoundation.org

Correspondence to:

colin.mitchell@phgfoundation.org

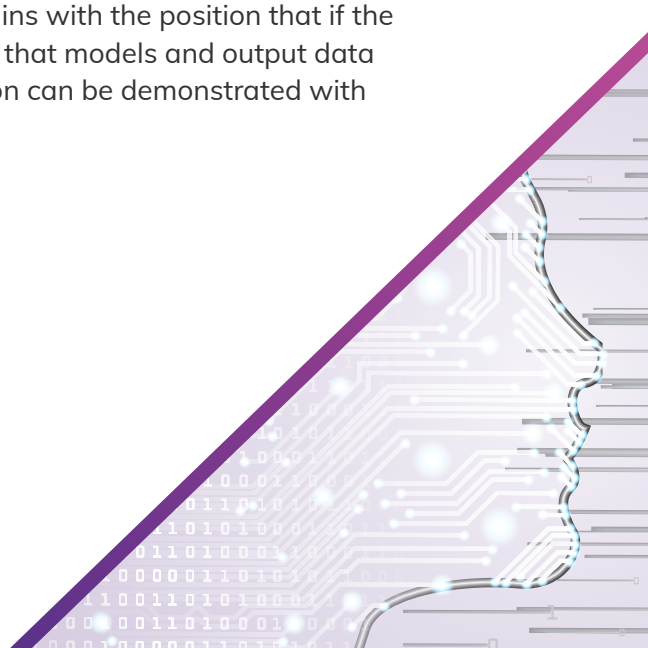
The PHG Foundation is a health policy think-tank and linked exempt charity of the University of Cambridge. We work to achieve better health through the responsible and evidence-based application of biomedical science. We are a registered company, no. 5823194.

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.

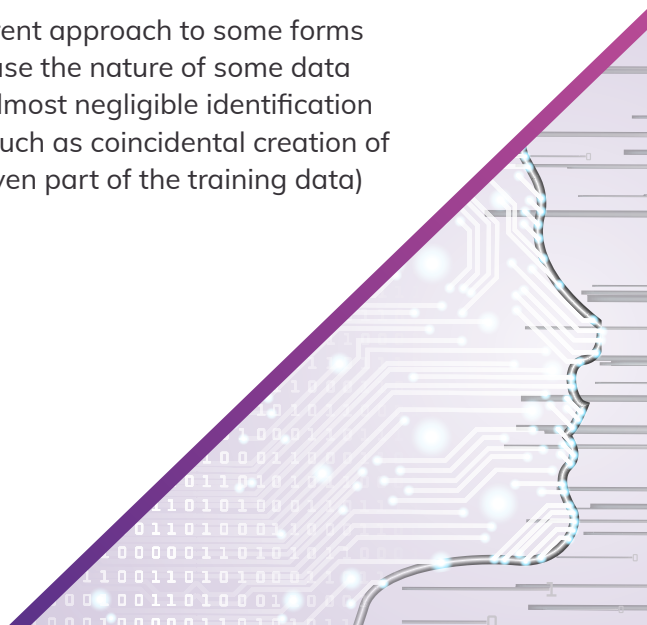


Unfortunately, demonstrating that data are no longer 'personal data' is not straightforward. Any assessment needs to be comprehensive and involves careful scrutiny of a wide range of factors including the nature of the data itself, the range of risks and attacks that could threaten the synthetic dataset 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)



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:

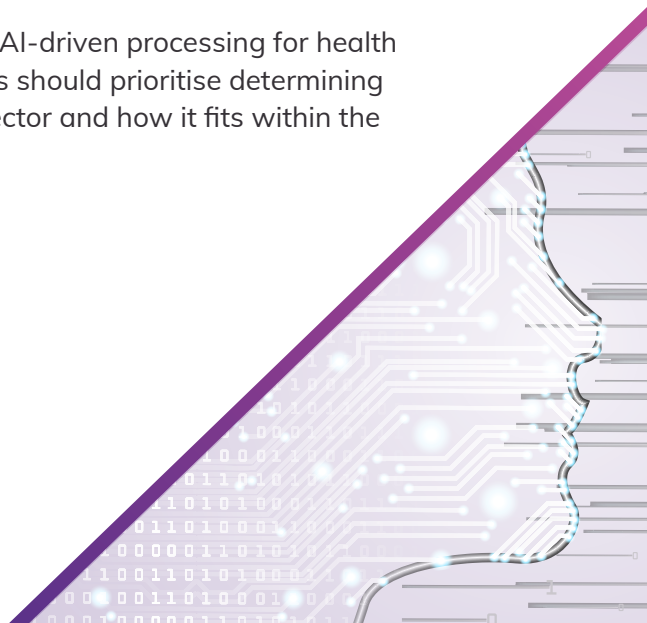
1. synthetic data developers and users should continue to follow best practice in relation to data protection impact assessments and anonymisation in assessing the identifiability and other data protection risks arising from processing.

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.



Contents

Introduction

- Methodology and approach

1. Introduction

Synthetic data can be thought of as artificial data that closely mimic the properties and relationships of real data. This concept is not new[1] but recent technological advances have seen a growth in attention on synthetic data in a range of sectors, including health and life sciences,[2] where there are a range of potential use cases including facilitating the development and validation of medical devices.

To this end, the Medicines and Healthcare products Regulatory Agency (MHRA) and the Clinical Practice Research Datalink (CPRD) have developed expertise and experience in the creation of synthetic datasets that can be used for training purposes or to improve algorithms.

In the context of patient health care data, a high fidelity or utility synthetic dataset captures complex clinical relationships and may be clinically indistinguishable from real patient data. There is a general assumption that high-utility synthetic data are associated with a higher privacy risk because they are closer to the real data. However, this may depend on the approach used to generate the dataset and the regulatory status of synthetic data is currently uncertain.[3]

While the generation of synthetic data begins with processing real patient data, it is unclear whether the resulting synthetic data remain private or subject to data protection law as 'personal data', or under what conditions this may be the case. In this report we seek to respond to this uncertainty and identify whether, or in what circumstance synthetic health data are 'personal data'.

This involves consideration of: the relevant technical approaches to, and resulting forms of, synthetic data in this context (Section 2); the case law and regulation relevant to the scope of 'personal data' in data protection law and (Section 3); and a synthesis of these elements to determine how data protection law may or should apply (or not apply) to relevant forms of synthetic data (Sections 4-5). We highlight key considerations for relevant stakeholders throughout or analysis and provide overall conclusions and recommendations in Section 6.

---

1 Rubin, D.B. (1993), Discussion: Statistical Disclosure Limitation, Journal of Official Statistics, 9(2), 461–468)

2 Campbell F, Hewitt W. Trends in digital health: Is synthetic data the real deal? Bristows Inquistivie Minds. 23 March 2023. Available at: <https://inquisitiveminds.bristows.com/post/102ib2d/trends-in-digital-health-is-synthetic-data-the-real-deal> accessed 27 February 2023

3 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2021 May;37(2):819-51; Stadler T, Oprisanu B, Troncoso C. Synthetic data–anonymisation groundhog day. In31st USENIX Security Symposium (USENIX Security 22) 2022 (pp. 1451-1468)

Data protection law in the UK is currently closely aligned with the EU General Data Protection Regulation (EU GDPR) and is governed by the UK's own version of the regulation, the UK GDPR. This diverges only slightly from the EU regulation to give effect to the UK's exit from the EU. Additionally, more specific rules and derogations are contained in the Data Protection Act (DPA) 2018.

However, the UK Government has set out its intention to reform data protection law and has introduced a Data Protection and Digital Information Bill containing some potentially relevant proposed changes. We consider this more fully in Section 3 but while further reform is subject to debate, this research is rooted in the principles and approaches developed over decades in European Union law and by the Council of Europe in sequential International Conventions, 108 and 108+, to which the UK remains subject. The primary legal focus of this research is therefore the UK GDPR and DPA 2018, with significant consideration of relevant EU jurisprudence and scholarship.

Methodology and approach

We aimed to evaluate whether and/or how data protection law applies to synthetic data as 'personal data' when generated from real patient data; to identify factors or circumstances which impact this assessment and; to develop recommendations for synthetic data developers, regulators and policymakers.

In order to deliver this project, the PHG Foundation drew on extensive in-house knowledge and experience in the context of data protection law and identifiability in the health data context.

The PHG Foundation team worked closely with CPRD/MHRA through the course of the research, in particular to develop an accurate understanding of the technical processes involved in the generation of synthetic data and the nature of the resulting datasets.

Objective

To evaluate whether or how data protection law applies to synthetic data as 'personal data' when generated from real patient data; to identify factors or circumstances which impact this assessment and; to develop recommendations for synthetic data generators, regulators and policymakers.

Project outline

The work included the following elements:

- understanding the context and the technical nature of synthetic data generation through review of documentation shared by the MHRA/CPRD team and through close liaison with MHRA and CPRD colleagues

- legal analysis of key elements of data protection law which impact the extent to which synthetic data may be considered personal data. Including:
  - review of legislation and case law in the UK and EU
  - review of guidance and recommendations from data protection authorities in the UK and EU
  - review of relevant literature on the scope of personal data, privacy and synthetic data in the health context.
- synthesis of the findings of the legal analysis with the technical context to identify conclusions and recommendations on the application of data protection law to synthetic data generated from real patient data.

Synthetic data

- What are synthetic data?
- Examples of synthetic patient data: CPRD and MHRA synthetic datasets
- Privacy and synthetic data
- Summary

2. Synthetic data

Synthetic data generation is a highly technical and rapidly developing field with a range of potential methods and approaches that can be adopted to develop new artificial data.

In this section we describe key approaches to synthetic data generation in the health context with a specific focus on examples developed by the CPRD and MHRA. We also highlight some of the additional safeguards and privacy measures that are commonly applied in synthetic data generation. This provides a foundation for our subsequent consideration of the legal status of synthetic data in Section 4, when viewed through the lens of the relevant law and interpretations considered Section 3.

What are synthetic data?

Synthetic data can be thought of as artificial data that closely mimic the properties and relationships of real or source data.[4] The degree to which the source data's correlations and relationships are preserved is dependent on the way in which the synthetic data is generated and for what purposes.

There are three general approaches for synthetic data generation. The first relies on the statistical properties of real data, such as population distributions (e.g. known prevalence of disease in groups and will require expert advice and knowledge of the relationships among such data, e.g. age to risk of cardiovascular disease, is relied upon in an attempt to mimic real data and ensure it is coherent.

While this approach is useful when real data are difficult to access, the complex relationships will be difficult to capture.[5] This could be generated from sampled data through mapping and replicating patterns in the data, or by using a combination of statistical information and rules based on expert knowledge to simulate patterns.[6]

---

4 Myles P, Ordish J, Branson R. Synthetic data and the innovation, assessment, and regulation of AI medical devices. RF Q. 2021;1:48-53

5 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2021 May;37(2):819-51; Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KS, Myles P, Granger D. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. arXiv preprint arXiv:1812.10404. 2018 Dec 21

6 Ibid

A second approach starts with the source data which is then manually obfuscated and manipulated in an attempt to closely preserve the relationships among the data points without revealing data subjects or other identifiers. Such an approach may be particularly useful when only part of the real world data needs to be generated, e.g., missing values need to be input via data imputation processes.[7]

A third approach uses machine learning techniques such as generative adversarial networks (GANs) or neural networks to decipher how the data points relate to each other. It is therefore possible that the AI model may find unusual correlations inherent in the data that may be missed by domain experts, therefore resulting in a truer approximation of the 'real' data. The sophistication of these models can recreate the authentic relationships between data points and outliers resulting in high fidelity synthetic data. This technique can be used to generate semi and fully synthetic data.[8] However, while machine learning can be very good at understanding such underlying complex relationships, black box algorithms may be too opaque for purposes where transparency regarding the underlying logic needs to be understood and trusted.[9] Transparent methods such as Bayesian networks are therefore important in generating synthetic health data from real patient data.[10]

In all three of these approaches the data may be manually reviewed for apparently incorrect correlations or data points, including to assess if any accidental or coincidental matches have been made to real living data subjects (where applicable). Such checks could include running rules or placing 'bounds' on the dataset to ensure that the values make biological sense e.g., no-one over the age of 150 is included in the dataset.[11] Expert analysis and oversight (medical and data scientists) are therefore potentially important in both the generation and evaluation stages of synthetic data.

The terms fidelity and utility are often used interchangeably to understand how 'clinically meaningful'[12] synthetic data is, and as previously mentioned can often be categorised on a scale from low to high fidelity/utility.[13] These terms refer to how well synthetic data maintains the underlying patterns and correlations in the source data and it is therefore likely that there is a relationship between fidelity/utility and privacy risks.

---

7 Ibid

8 Ibid

9 Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KS, Myles P, Granger D. Machine learning and AI research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. arXiv preprint arXiv:1812.10404. 2018 Dec 21

10 Ibid

11 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2021 May;37(2):819-51

12 Ibid

13 See for examples: Dankar FK, Ibrahim M. Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences. 2021 Feb 28;11(5):2158; Amor E. Generate high-fidelity synthetic data with the Statice SDK. Statice Blog 2021. Available at: <https://www.statice.ai/post/generate-high-fidelity-synthetic-data-with-the-statice-sdk> accessed 24 January 2023

As we discuss more fully in Section 4, synthetic data generation is commonly viewed as a form of anonymisation or as a privacy-enhancing technology (PETs).[14] However, while synthetic data may provide greater opportunity for enhanced privacy over processing of real data, this is not always the primary driver for its generation. It is true that some use cases for synthetic data are based on reducing the burden of legal compliance (and risk) that would be in place for sharing and processing real patient data.[15] However, synthetic data can also be useful for gap-filling missing information needed to test products, software or sections of code,[16] enabling cost-effective alternatives to generating suitable real data to train or validate AI tools as well as a method for discovering bias in real world data or identifying new relationships among data points.[17]

Examples of synthetic patient data: CPRD and MHRA synthetic datasets

CPRD has generated a number of synthetic datasets that can be used for training purposes or to improve algorithms or machine learning workflows. Two are high fidelity datasets for cardiovascular disease and COVID-19. Both were generated from de-identified primary care data, extracted from the CPRD Aurum database.[18] CPRD has also developed a medium fidelity dataset, the CPRD Aurum Sample dataset which resembles the real world CPRD Aurum.[19] These datasets were generated using a synthetic data generation and evaluation framework,[20] developed under a grant from the Regulators' Pioneer Fund launched by the formerly known Department for Business, Energy and Industrial Strategy (BEIS) (now the Department for Science Innovation and Technology (DSIT)) and managed by Innovate UK. This generation and evaluation framework, along with the datasets themselves are owned by the Medicines and

---

14 Information Commissioner's Office (ICO). Data anonymisation, pseudonymisation and privacy enhancing technologies guidance (Draft Guidance, September 2022). Available at: <https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/> accessed 26 January 2023

15 UK Statistics Authority. Ethical considerations relating to the creation and use of synthetic data: Synthetic data and ethics. 19 October 2022. Available at: <https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/pages/2/> accessed 26 January 2023

16 Ibid

17 Ibid

18 CPRD, 'Synthetic Data' (12th January 2023) Available at: <https://cprd.com/synthetic-data#CPRD%20cardiovascular%20disease%20synthetic%20dataset> accessed 12th January 2023

19 Ibid

20 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2021 May;37(2):819-51

Healthcare products Regulatory Agency.

CPRD state that their 'high fidelity synthetic datasets replicate the complex clinical relationships in real primary care data while protecting patient privacy as they are wholly synthetic.'[21] As such, they can be appropriately used for complex statistical analyses as well as machine learning and artificial intelligence (AI) research applications. CPRD, MHRA and academic partners at Brunel have been at the forefront of producing papers outlining the generation and evaluation process and have attempted to create a framework to assess privacy protection and accuracy to the ground truth data.[22]

Privacy and synthetic data

There is a finely balanced trade-off between preserving data subject privacy and utility/fidelity in synthetic data. In some cases, synthetic data may be considered to require additional privacy measures to safeguard against identification or the drawing of inferences about an individual from the data. One such method is called differential privacy. Differential privacy obscures by adding a certain amount of noise. The additional noise will mean that the output data's accuracy cannot be entirely trusted and there may be uncertainty about whether the answer shown is a result of the algorithm adding noise or not. Methods such as Laplace distribution can be applied to add noise to a larger range of data to increase the level of anonymity.[23] Such technology may enable sensitive data to be released to the public while protecting the anonymised data from a range of attacks.

However, there are some limitations. Differential privacy works best on large datasets and is less suitable to smaller datasets due to their larger distributive range. Moreover, an estimation from repeated queries may reveal the ground truth, also known as composition theorem. Consequently, a fine balance has to be struck between further increasing noise (at the risk of reducing its utility) against a successful privacy attack. To tackle this, privacy accounting is one useful method where a maximum privacy loss limit can be enforced by a data curator, i.e., a privacy budget. In such cases, there is a maximum limit on how many queries can be made.

---

21 CPRD, 'Synthetic Data' (12th January 2023) Available at: <https://cprd.com/synthetic-data> accessed 24 January 2023

22 Wang Z, Myles P, Tucker A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence. 2021 May;37(2):819-51; Myles P, Ordish J, Branson R. Synthetic data and the innovation, assessment, and regulation of AI medical devices. RF Q. 2021;1:48-53'; Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ digital medicine. 2020 Nov 9;3(1):1-3

23 See Roth and Dwork's paper for more detailed information on how Laplace distribution works. Dwork R. Dwork C, Roth A. The algorithmic foundations of differential privacy, Foundations and Trends in Theoretical Computer Science. 2014;9(3-4):211-407

Ultimately, the privacy strength and usefulness of any given dataset is highly context dependent and a range of factors (discussed more fully in Section 4) will be relevant. These include method of generation, whether further data privacy measures have been used to mitigate or prevent attacks (differential privacy or k-anonymity etc), how datasets are maintained and how security is safeguarded.

Summary

Synthetic data can generally be defined as artificial data that closely mimics the properties and relationships of real or source data. However, within this broad definition there are a range of 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.

Some use cases for synthetic data are based on reducing the burden of legal compliance (and risk) that would be in place for sharing and processing real 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.

The identifiability and privacy implications of any synthetic dataset or synthetic data model will be highly context dependent. A range of factors are likely to play a part including the methodology used and whether further data privacy measures have been used to mitigate or prevent attacks, such as differential privacy.

We now turn to the legal framework, identifying the relevant definitions and interpretations that will have a bearing on whether, or in what circumstances synthetic data may constitute personal data.

The Data Protection Law framework

- Data protection law in the UK and EU
- Material scope of data protection law
- Summary

3. The Data Protection Law framework

An assessment of whether synthetic health data constitute 'personal data' or fall within the 'material scope' of the law requires a careful consideration of the relevant legal definitions and their interpretations. This particularly includes the interpretation of identifiability and anonymisation within data protection law. In this section we set out the components that make up the material scope of the current UK GDPR and how these have been approached and interpreted by courts and in authoritative guidance.

The UK law is currently closely aligned with the EU GDPR and decades of EU precedent, guidance and scholarship remain highly relevant when interpreting and applying the concept of 'personal data'. The UK government has recently introduced proposals for reform which contain some minimal changes that could impact material scope.

However, we identify strong reasons to believe that EU definitions and interpretations will remain highly relevant to understanding the scope of 'personal data' in UK law. Therefore, the focus in this report will remain on the interpretation of 'personal data' in current law. We begin this section by situating the UK legal position in the wider international context.

Data protection law in the UK and EU

The current UK legal framework for data protection is derived from the EU General Data Protection Regulation which currently applies in virtually the same form in the United Kingdom as the 'UK GDPR'. This sits alongside the Data Protection Act 2018 which tailors and supplements some parts of the general regulation. The whole framework is overseen by the independent authority, the Information Commissioner's Office (ICO) which provides guidance on the application of the law and is responsible for handling complaints, with the discretion to levy fines and carry out enforcement where there has been a breach of data protection law. The courts will also hear claims of a breach of data protection law.

The GDPR has a very close relationship with human rights instruments, in particular those issued by the Council of Europe; the international organisation which is distinct from the European Union, and responsible for the European Convention on Human Rights. The Council of Europe adopted the first international instrument on data protection in 1981 to set standards and principles to ensure respect for the fundamental right of all individuals 'with regard to processing of personal data'.[24]

---

24 Council of Europe, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No.108, 1981)

Convention 108, as it is known, aims to promote the respect for privacy and protection of personal data at a global level, and which fifty-five countries have now either ratified or signed. Convention 108 contains the seeds of the EU's data protection principles, including the core principles of lawful, fair, and purpose-limited processing, and it was highly influential in the EU's development of its own Data Protection Directive (DPD) in the 1990s. However, unlike the Council of Europe, the EU was not a fundamental rights organisation and its Directive was primarily designed to harmonise national rules on data protection and ensure the free movement of personal data within its economic area. In fact, the treaty basis for the DPD was a provision that aimed to harmonise the 'internal market' of the European Community.[25]

To remove obstacles to the free flow of data, the Directive also aimed to ensure that the level of protection of fundamental rights, in particular the right to privacy, were consistent across the EU (recital 7). However, as a Directive, it left the means of implementation up to individual Member States and also provided a considerable margin of appreciation as to the level of, and nature of, the protection of personal data in each State.

In the decades following the adoption of the Directive, technological and social changes coincided with a shift in the status of data protection in the EU. As the EU developed and expanded, it adopted a Charter of Fundamental Rights, including a 'fundamental right to the protection of personal data'.[26] This was also incorporated in one of the EU's primary treaties alongside a new legal basis for the EU, to legislate to establish rules relating to the processing of personal data, separate to the need to harmonise an internal market. This became the foundation for the GDPR, which was inspired by a need to 'put individuals in control of their personal data' and the strong statement that: '... individuals have the right to enjoy effective control over their personal information'.[27] Data protection is a fundamental right in Europe, enshrined in Article 8 of the Charter of Fundamental Rights of the European Union, as well as in Article 16(1) of the Treaty on the Functioning of the European Union (TFEU), and needs to be protected accordingly.'[28]

Now that the UK has left the EU, the UK GDPR sits awkwardly as a set of rules and norms which derive from a fundamental right to data protection which is not recognised explicitly in UK law. The UK does recognise a right to respect for private and family life under Article 8 of the European Convention on Human Rights (ECHR) and requires this right to be given direct effect through the courts and by lawmakers via the Human Rights Act 1998.

---

25 Consolidated Version of the Treaty on Functioning of the European Union [2016] OJ C202/95, Art 115

26 Charter of the Fundamental Rights of the European Union [2012] OJ C326/391, Article 8(1)

27 Ibid

28 European Commission. Safeguarding Privacy in a Connected World: A European Data Protection Framework for the 21st Century. (Communication) COM (2012) 9 final, 2-5

However, the influence of the separate right to data protection (i.e., privacy) recognised in the EU Charter on interpretations of EU data protection law could lead to a divergence between the EU and the UK, driven by differing interpretations of the same provisions and concepts.[29]

Current UK Data Protection Law and proposals for reform

While the UK was a member of the European Union, the process of replacing the 1995 Data Protection Directive resulted in the development of the EU General Data Protection Regulation (2016/679) (the "EU GDPR"). This came into force in, and applied directly in the UK, from the 25th May 2018, supplemented by the UK Data Protection Act (DPA 2018) (in particular Part 2 of the Act) which exercised derogations provided by the EU GDPR and allowed aspects of the law to be tailored by each Member State.

The EU GDPR was incorporated into UK law in December 2020, at the end of the EU Transition Period, under Section 3 of the European Union (Withdrawal) Act 2018 (EUWA 2018). Minor modifications were made by the Data Protection, Privacy and Electronic Communication (Amendments etc) (EU Exit) Regulations 2019 under the power in Section 8 EUWA 2018 to create the UK GDPR.

The UK GDPR and the DPA 2018 therefore provide the core of data protection law across the entire United Kingdom (different laws apply to law enforcement and intelligence services processing – outside the scope of this work) with provisions that are nearly identical to those across the EU. However, the UK Government has set out its intention to reform data protection law, first introducing the Data Protection and Digital Information (DPDI) Bill in Summer 2022 and then (following changes in Government) in an updated version of the DPDI Bill (No.2) in March 2023.

The Data Protection and Digital Information (No.2) Bill

The clear intention behind the DPDI (No.2) Bill is to simplify the UK's data protection framework and reduce the burdens on organisations, while maintaining high data protection standards.[30] This ambition extends to adjusting the definition and approach to be applied to determining when whether data are 'personal data'.

Clause 1 of the Bill proposes amendments to the definition of personal data by expanding on when an individual may be identified or identifiable. It sets out a new Section 3A to be inserted in the Data Protection Act 2018 which provides for two cases in which information being processed by a controller or processor counts as information relating to an identifiable individual and is therefore personal data for the purposes

---

29 Note that Article 29 Working Party Guidance, Opinion 4/2007 states that the right to family and private life are recognised as two separate arms, i.e., family life and privacy, under EU law. Whether this is the case in UK law is left open for interpretation

30 See explanatory notes to the Bill. para1: Available at: <https://publications.parliament.uk/pa/bills/cbill/58-03/0265/en/220265env2.pdf> accessed 16 March 2023

of the legislation: The first case is where the controller or processor can themselves identify a living individual from the information they are processing by reasonable means, which would encompass for example the re-identification of pseudonymised or coded data by the controller if they are in possession of the relevant key or code.

The second case is where the controller or processor knows, or ought reasonably to know, that another person is likely to obtain the information as a result of the processing - for example, somebody with whom the information is shared - could identify a living individual by reasonable means.

The specific proposed wording is set out below.

"3A Information relating to an identifiable living individual

(1) for the purposes of this Act, information being processed is information relating to an identifiable living individual only in cases described in subsections (2) and (3).

(2) the first case is where the living individual is identifiable (as described in section 3(3)) by the controller or processor by reasonable means at the time of the processing.

(3) the second case is where the controller or processor knows, or ought reasonably to know, that—

(a) another person will, or is likely to, obtain the information as a result of the processing, and

(b) the living individual will be, or is likely to be, identifiable (as described in section 3(3)) by that person by reasonable means at the time of the processing.

(4) the reference in subsection (3)(a) to obtaining the information as a result of the processing includes obtaining the information as a result of the controller or processor carrying out the processing without implementing appropriate technical and organisational measures to mitigate the risk of the information being obtained by persons with whom the controller or processor does not intend to share the information.

(5) for the purposes of this section, an individual is identifiable by a person "by reasonable means" if the individual is identifiable by the person by any means that the person is reasonably likely to use.

(6) for the purposes of subsection (5), whether a person is reasonably likely to use a means of identifying an individual is to be determined taking into account, among other things—

(a) the time, effort and costs involved in identifying the individual by that means, and

(b) the technology and other resources available to the person."

While this proposal is to amend the DPA 2018 and insert a new section in that Act, another amendment would apply to Section 3A of the UK GDPR so there can be no doubt that it is intended to amend the definition of 'personal data' in the Regulation.

The impact of these proposed changes is currently unclear. To an extent, these proposals are to an extent minor amendments or clarifications of current GDPR definitions and concepts and neither the Government (in explanatory notes) nor legal commentators suggest that these are significant changes to the scope of personal data. Given these proposals are at an early legislative stage—and are highly likely to change during the legislative process—we will focus on the current definition and interpretations of the law in place at the time of writing, although there are several elements which could foreseeably adjust in meaningful ways when data are considered 'personal data' or otherwise non-personal/anonymised under UK law in the future.

Material scope of data protection law

The UK GDPR applies only to the 'processing' of 'personal data'. Processing is a catch all term that describes almost anything that may be done with data, including storage.[31]

The UK GDPR defines personal data under Article 4(1) as:

> 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.'

There are several components to this definition that need to be unpacked in order to assess whether synthetic data may constitute personal data. Some of these are relatively straightforward whereas others are not.

The individual elements of the definition of personal data are not expanded on in the main text of the law but they have been interpreted and refined over time in EU case law and in guidance from authoritative EU data protection bodies, the Article 29 Working Party (WP29) and (its replacement under the GDPR) the European Data Protection Board (EDPB). In particular, the WP29 Opinion 4/2007 on the concept of personal data provides a benchmark interpretation drawing on relevant case law, which is echoed in guidance from data protection authorities including, in the UK, the ICO.[32] Subsequent case law and commentary add additional support in unpacking these components.

---

31 Article 4(2), UK GDPR

32 Article 29 Working Party Guidance, Opinion 4/2007

'Any information'

The first component is 'any information'. This phrase evidently entails a broad interpretation. Article 29 Guidance (WP 29) Opinion 4/2007 suggests that it encompasses any statement that could be made about a person, both objective (e.g., blood type) and subjective (opinions or assessments). No limitation exists that the information must be 'true'; this is reflected in the fact that data subjects have a right to rectification of erroneous data is held about them under Articles 16, 18 and 19 GDPR, including under the GDPR's principle of accuracy (Article 5(1)(d)). It includes data concerning the individual's private and family life and information on their activities such as working relations, economic, social, mental etc., and consequently, encompasses all information 'regardless of their capacity or position,'[33] relating to an individual. To demonstrate how expansive this can be, Opinion 04/2007 gives the example of prescription information still amounting to personal data relating to the physician prescribing the medication, even where the patient is anonymous as it reflects an identifiable person's activities. Additionally, there is no limitation on how the information is presented e.g., paper versus binary code (among other formats such as sound and imagery).

For synthetic patient or health data generation, the form of records processed and generated would clearly constitute information within the material scope of the UK GDPR.

'Relating to'

'Relating to' is a key part of establishing if data is personal. The interpretation of 'relating to' is important for understanding what information is sufficiently linked to a natural person to amount to their personal data.[34] However, the concept and its interpretation demonstrates the wide ambit that data protection law intends to draw. In many cases the content of the information will be clearly 'about' an individual, for example medical test results are clearly about the person tested. However, other information may not obviously be about a person but they could be used to, or result in an, impact on their rights and interests. According to WP29 and CJEU, this requires a focus on the 'content', 'purpose' and 'effect' of the data,[35] to determine whether it is linked to a particular person.

Content is given its ordinary interpretation to mean it is 'about' someone, to be assessed in light of the case.'[36] The purpose element 'can be considered to exist when the data

33 Ibid, p.7

34 Ibid

35 Ibid, 10. Case C-434/16 *Peter Nowak v Data Protection Commissioner* [2017] ECR I-994, [35]

36 The Information Commissioner's Office. What is the meaning of 'relates to'?. Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/what-is-the-meaning-of-relates-to/> accessed 27 February 2023

are used or are likely to be used... with the purpose to evaluate, treat in a certain way or influence the status or behaviour of an individual.'[37] Finally, the result element is present when, regardless of whether the content or purpose elements are present, 'their use is likely to have an impact on a certain person's rights or interests.'[38] There is no requirement that the result have a major impact, the impact will be sufficient where the individual is 'treated differently from other persons as a result of processing such data.'[39]

The content, purpose or result elements are 'alternative conditions' and are not to be considered 'cumulative'.[40] There is also nothing to suggest that each element may relate to different individuals at the same time.[41] Whilst some forms of data are clearly linked to an individual, e.g., certain forms of health data, others are not so clear cut but may still nevertheless impact the rights and protections afforded to individuals under the Data Protection Act 2018 when linked with other available data.[42] Equally, it is possible for information to identify an individual but not in fact be found to 'relate to' them (for example, if someone's name is incidentally included in correspondence about someone else).[43]

Synthetic patient data would be most likely to be considered to 'relate to' an individual in terms of the content being about a person if an individual is identified or identifiable (see below). However, this may not be as straightforward where synthetic patient data is not quite identical to a real patient record.

'Natural person'

A natural person is a living individual. Under the Universal Declaration of Human Rights 'everyone has the right to recognition everywhere as a person before the law' where the only limitation is species i.e., being human, this seemingly encompasses all human beings. However, some qualifiers have been added elsewhere in domestic law and may therefore be different across different jurisdictions. For example, the ICO adds[44] that a natural person means a living human being as opposed to legal entities such as companies who are recognised legal persons.[45] Additionally, Recital 27 makes it clear that the GDPR does not apply to deceased persons.[46]

---

37 Article 29 Working Party Guidance, Opinion 4/2007, p.10

38 Ibid

39 Ibid

40 Ibid, p.11

41 Ibid

42 PHG Foundation. The GDPR and genomic data: The impact of the GDPR and DPA 2018 on genomic healthcare and research. PHG Foundation, May 2020. Available at: <https://www.phgfoundation.org/report/the-gdpr-and-genomic-data> accessed 27 February 2023

43 The Information Commissioner's Office. What is the meaning of 'relates to'? Available at: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-is-the-meaning-of-relates-to/> accessed 27 February 2023

44 Information Commissioner's Office (ICO), 'Guide to Data Protection: What is personal data?' Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/> accessed 27 February 2023

45 Legal persons are non-human entities capable of holding legal personality i.e., to sue and be sued. Companies are a good example of this

46 General Data Protection Regulation (GDPR), Recital 27

As the recital itself states, that does not mean that Member States or the UK do not have other rules regulating the use of deceased persons personal data, much like anonymous data which although it falls outside the Regulation, its processing may still be subject to other legal rules. Consequently, there may be practical issues for controllers identifying whether data subjects are still alive. Additionally, living individuals have interests in other deceased persons' data, this may be particularly relevant for certain health data such as genomic information. Therefore, where it is not known if the data subject is alive it may be easier to treat all data subjects within a dataset as if they are living.

It is unclear from reading Article 4 if in utero human beings are out of the scope of the GDPR. The GDPR does not itself further define a 'natural person' and the ICO seems to only rule out non-natural persons from the definition of personhood for the purposes of data rights.[47]

Under domestic law, the UK has long held that *in utero* humans are not legal persons but whether they are natural persons, capable of being data subjects for the purposes of the GDPR depends on a loosely defined understanding of 'natural persons'.[48] This may have future relevance to synthetic data due to a possibility of creating a synthetic data profile that in future matches a no longer hypothetical human being. How the GDPR would approach such a challenge is currently unclear. Data relating to born children are, in practice, regulated more strictly by the ICO due to their vulnerability.[49]

'Identified or identifiable'

WP 29 Opinion 4/2007 states that a natural person can be considered identified when they can be distinguished from the rest of the group. Identifiable broadens this to encompass the possibility that an individual may be identified. This therefore means that the individual need not have been already identified for the data in question to be personal. Recital 26 UK GDPR offers further aid in interpreting identifiability by suggesting that for an individual to be considered identifiable, 'account should be taken of all the means reasonably likely used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.'

---

47 Information Commissioner's Office (ICO), 'Guide to Data Protection: What is personal data?' Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/what-is-personal-data/> accessed 6 March 2023

48 *Rance v Mid-Downs Health Authority* (1991) 1 All ER 801 (QB); *AG Ref No 3 of 1994* (1997) 3 All ER 936 (HL)

49 General Data Protection Regulation (GDPR), Recital 38 states 'children require specific protection with regard to their personal data as they may be less aware of the risks, consequences and safeguards concerned and their rights in relation to the processing of personal data...' See also specific guidance on how to navigate the processing of a child's personal data: Information Commissioner's Office (ICO), 'Children and the UK GDPR: What should our general approach to processing children's personal data be?'. Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/children-and-the-uk-gdpr/what-should-our-general-approach-to-processing-children-s-personal-data-be/> accessed 24 February 2023

The standard for determining whether information is identifying is whether there is a reasonable likelihood of identification taking into account all the circumstances and nature of the data at hand.[50] In general, this has been interpreted broadly, so that de-identified data (for example, stripped of any name, date of birth and other directly identifying attributes) is likely to be 'personal data' if it is still individual-level (as opposed to aggregated) data.

Reasonableness was further expanded on by the Explanatory Report to Modernised Convention 108 as, not 'requir[ing] excessively complex, long and costly operations.'[51] Moreover, Recital 26 GDPR states that, 'to ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of processing and technological developments.'[52] As such, assessing identifiability will require assessment over the life-time of the data; the longer it is stored, the greater the risk as technology advances.

This element is likely to be key in determining whether synthetic data can be said to be 'personal data' by virtue of identifying or potentially identifying an individual. Such identification maybe direct or indirect.

'Directly or indirectly'

Direct identification generally refers to the ability to identify individuals solely based on identifiers that are present in the data. For this reason, it is common practice across healthcare and research to remove all obvious identifiers such as names and individual patient numbers when de-identifying data.

A natural person is identifiable when they can be distinguished from all other members of the group.[53] This does not mean that the individual must already have been singled out, just the possibility of it means that they are identifiable, setting a high threshold.[54] Identifiers are information which hold a close relationship to an individual.

Article 4(1) refers to examples 'such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.' The meaning of directly or indirectly relates to the type of identifier. For example, National Insurance numbers are considered directly identifying as they are unique to one natural person, whereas an address could create an indirect link.

50 Recital 26 GDPR

51 Council of Europe, Explanatory Report 108, 2018, para 17. Note its first iteration (Explanatory Report 108, 1981, para 28) states that an identifiable person, 'means a person who is easily identified: it does not cover identification of persons by means of very sophisticated methods.'

52 Recital 26 GDPR

53 Article 29 Working Party Guidance, Opinion 4/2007, p.21-22

54 Ibid, p.12

The Explanatory Memorandum of the OECD Privacy Guidelines gives similar examples for the interpretation of direct and indirect identifiers/ linkages.[55] Consequently, direct identification usually relies on identifiers present in the data, whereas indirect identification requires other available data.[56] Nevertheless, whether a piece of information is identifying or not, in and of itself, will always be dependent on context e.g., some surnames are more common than others and therefore may be less likely to single out a specific individual than less common names.

However, given the increasing amount of publicly available data on individuals, indirect identification presents a challenge and ever-moving goal post for data controllers. Additionally, in the context of health data, Recital 35 notes that, 'personal data concerning health should include all data pertaining to the health status of a data subject which reveal information relating to the past, current or future physical or mental health status of the subject.'[57]

This is the same for genetic data which amounts to personal data where inherited or acquired characteristics of a natural person can be read through analysis of a biological sample from a natural person in question.'[58]

For synthetic health data of the nature described in the previous section, direct identifiers will almost certainly have been removed from patient data prior to processing. The crucial question is therefore whether an individual may be indirectly identifiable from the data (and other available sources). The text of the Regulation does not provide guidance on how to make this assessment directly but the recitals do set out the approach that should be taken.

'Means to identify'

The approach that should be taken to assessing identifiability is explained in recital 26 of the GDPR:

> 'To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.'

---

55 See OECD Privacy Guidelines 2013 at p.52 where they give civil registration numbers as a direct linkage and addresses as an indirect example

56 PHG Foundation, The GDPR and genomic data: The impact of the GDPR and DPA 2018 on genomic healthcare and research (PHG Foundation, May 2020). Available at: <https://www.phgfoundation.org/report/the-gdpr-and-genomic-data> accessed 27 February 2023

57 Recital 35 GDPR

58 Recital 34 GDPR

There must be a determination of whether there are 'means reasonably likely to be used', either by the controller or by another person. As part of this, 'account should be taken of all objective factors' required for identification. These include the technology that may be available to connect data or make inferences and a range of contextual factors that have been highlighted as relevant to an assessment of whether data are identifiable. The standard of reasonable likelihood applies to the means which could be used but this also sets a general benchmark for assessing identifiability. We explore this more thoroughly in Section 3 below.

Determining when means to identify are reasonably likely to be used is not straightforward. Case law and guidance help to set the interpretative limits to this concept and demonstrate when identification is not reasonably likely. For example, WP29 have referred to 'mere hypothetical possibility',[59]and the UK courts to a 'remote' chance[60] of identification, which would be insufficient. However, more recent CJEU case law has also suggested that a low level of identifiability risk could result in data being classified as personal.

In the case of Breyer, the CJEU adopted an expansive interpretation that suggested that not all information needs to be held by one person for it to be personal data.[61] In Breyer, the existence of legal channels allowing online media service providers to request that German State authorities obtain information from an internet service provider to identify a person from their dynamic IP address, constituted means that may be reasonably likely to be used to identify a data subject.[62] The fact that such means would only be used in exceptional circumstances such as cyberattacks was not a decisive factor.[63] The court set an apparently high standard for de-identification, ruling that such data would not be personal data only if 'the risk appears in reality to be insignificant' or if it was practically impossible to identify an individual.[64] However, it is worth noting that this ruling has been treated cautiously in an English court where it has been viewed as a decision that hinged on aspects of the German legal system.[65]

---

59 Article 29 Data Protection Working Party. Opinion 4/2007 on the concept of personal data. 2007, 1-26

60 *Department of Health v Information Commissioner* [2011] EWHC 1430 (Admin), [2011] WL 1151213

61 Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECR I-769

62 PHG Foundation, The GDPR and genomic data: The impact of the GDPR and DPA 2018 on genomic healthcare and research (PHG Foundation, May 2020). Available at: <https://www.phgfoundation.org/report/the-gdpr-and-genomic-data> accessed 27 February 2023

63 Ibid

64 Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECR I-769, para 49

65 *Mircom International Content Management and Consulting v Virgin Media* [2019] EWHC 1827 (Ch), [2019] 7 WLUK 245

Recital 26 also makes clear that all means are included and that it is not limited to certain categories of people who may access or seek to access this information. Opinion 04/2007 suggests that factors such as intended purpose, the structure of processing, the advantage expected by the controller, interests at stake for the individual, risk of breach (organisationally or technically) should be considered.[66] Moreover, such factors should be considered in light of the state of the art in technology at the time.[67] Therefore, the longer the data is intended to be stored and used will increase the likelihood that such data amounts to personal data.

The lifetime of the data is therefore an important consideration, including for synthetic data. This can be mitigated through developments and system updates to ensure the security of the data is maintained.

There is no doubt that the CJEU and EU data protection bodies have adopted an expansive approach to the scope of 'personal data' over the years, leading scholars to consider the extent to which almost any information could be considered personal data.[68] However, this is a risk based approach which provides scope for some data to be considered non-personal or 'anonymous' data. Although the GDPR does not define anonymous data, aspects of the law and guidance help to delineate the outer limits of personal data.

Anonymised and pseudonymised data

Anonymous information is referred to in recital 26 as 'information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.' However, the WP 29 in a 2014 report on Anonymisation Techniques emphasised that true anonymisation is 'irreversible',[69] setting a potentially very high threshold for data to be considered anonymous. However, this interpretation met with considerable academic criticism for its increasing impracticality,[70] and authorities, including the UK ICO have more practical standards which account for a small residual risk of identification that may remain while still satisfying the threshold of anonymity (see further Section 4 below).

66 Article 29 Working Party Guidance, Opinion 4/2007, p.15

67 Ibid

68 Purtova, N. (2018). The law of everything . Broad concept of personal data and future of EU data protection law. 9961. Available at: <https://doi.org/10.1080/17579961.2018.1452176> accessed 27 February 2023

69 Article 29 Working Party Guidance, Opinion 05/2014, p.5

70 See for example: Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. UCLA l. Rev.. 2009;57:1701 at 1742 and 1759; Schwartz PM, Solove DJ. The PII problem: Privacy and a new concept of personally identifiable information. NYUL rev. 2011;86:1814

Some confusion was also caused by the express reference to a category of data which have undergone 'pseudonymisation' explicitly in recital 26 GDPR as 'personal data'. This is because in some jurisdictions and sectors, pseudonymous data- a process where real-world identifiers are removed from data and replaced with a key, cipher or code so that an individual cannot be easily identified from the data without that key or code- had been considered to be a form of anonymisation. This led to significant uncertainty about whether such data are always considered personal data- i.e. whether a key has to be deleted in order to anonymise the data or whether it is feasible to safeguard and separate the key sufficiently for it to be considered anonymous.[71]

In previous work we argued strongly that the risk based definition of personal data must logically mean it is feasible to sufficiently safeguard the key from the semantic information to fall outside the scope of personal data.[72] In part the answer depends whether an 'absolute' or 'relative' approach is adopted; i.e. whether identifiability is judged according to the abilities of anyone and everyone (including the data controller) to re-identify data or whether the relevant question is whether the data are 'identifiable' in the 'hands' of a specific actor.

The CJEU in the aforementioned case of Breyer made a decisive contribution by adopting the relative approach to assessing identifiability, which strongly supports the conclusion that pseudonymous data do not always remain personal data despite the level of separation and safeguards applied to protect them from re-identification.[73] This is also the approach that the ICO has adopted in its latest draft guidance: 'you may be able to disclose a pseudonymised dataset (without the separate identifiers) on the basis that it is effectively anonymised from the recipient's perspective.'[74]

The status of pseudonymised information is important in the consideration of synthetic health data because it is likely that identifiers will have been removed from both input or training data and from output synthetic datasets. However, if it can be argued that the output data is reasonably likely to be capable of being recombined with identifying information this would only constitute 'pseudonymised' data and thereby fall within the scope of personal data.

---

71 Rumbold JM, Pierscionek B. The effect of the general data protection regulation on medical research. Journal of medical Internet research. 2017 Feb 24;19(2):e47. Shabani M, Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation. European Journal of Human Genetics. 2018 Feb;26(2):149-56. Donnelly M, McDonagh M. Health research, consent and the GDPR exemption. European journal of health law. 2019 Apr 24;26(2):97-119. Quinn P. The anonymisation of research data—a pyric victory for privacy that should not be pushed too hard by the EU data protection framework?. European Journal of Health Law. 2017 Oct 19;24(4):347-67

72 Mitchell C, Ordish J, Johnson E, Brigden T, Hall A. The GDPR and Genomic Data: PHG Foundation report on the impact of the General Data Protection Regulation and Data Protection Act 2018 on Regulating Genomic Technologies in Healthcare for the Information Commissioner's Office. PHG Foundation. 2020 Available at: <https://www.phgfoundation.org/report/the-gdpr-and-genomic-data> accessed 27 February 2023

73 Mourby M, Mackey E, Elliot M, et al. Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law & Security Review. 2018; 34(2): 222-223

74 ICO. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Chapter 3: pseudonymisation. February 2022

Summary

At present UK data protection law in the form of the UK GDPR is almost identical to the underlying EU regulation. However, following the UK's exit from the EU, there is scope for divergence and the UK Government has set out proposals for reform of elements of the law in a recent Data Protection and Digital Information Bill. This includes some changes to the definition of 'personal data' which is the crucial element in this research. However, at present these changes are uncertain and they are not recognised to be significant in the explanatory notes or commentary so far.

Combined with other reasons that UK law is likely to remain fairly closely aligned with EU regulation in this area (the presence of an underlying international convention and the practical implications of diverging too far from EU law), in this research we examine the material scope of the law as it is currently defined, in light of existing EU case law and guidance.

There are multiple components to the material scope of the UK GDPR. For synthetic health data most of these are likely to be easily met, the greatest challenge and ambiguity lies in determining whether information can be said to relate to an identified or identifiable living individual, directly or indirectly. This will involve an assessment of whether means are reasonably likely to be available to identify an individual from synthetic data according to the state of the art at the time.

A risk based and contextual assessment is required to assess whether synthetic data may be identifying, both at the time of processing and as technology develops over time. Although there are indications that even a relatively low risk of identification may suffice for synthetic data to be considered personal data, the latest approach of the CJEU and ICO provide room for determining that synthetic data could be considered anonymous or non-personal data.

Synthetic data as ‘personal data’?

- Relevant law, guidance and commentary
- Current Data Protection authorities’ approaches to synthetic data
- Legal scholarship and commentary on synthetic data
- Summary

4. Synthetic data as 'personal data'?

The central question in this research is whether or not synthetic health data are considered 'personal data', or, in what circumstances this may be the case. We set out the relevant aspects of the law in the previous section, in this section we consider the available specific case law, authoritative guidance and legal commentary on synthetic data and how these questions are likely to be answered by courts and regulators.

On this basis, we set out a range of the general and specific factors that are likely to be relevant to determining whether synthetic health data are personal data, highlighting a potential distinction between a current 'orthodox' approach and potential alternative ways of addressing this question in Section 5. We highlight key considerations for relevant stakeholders where they arise in this analysis.

Relevant law, guidance and commentary

Our analysis of relevant law, authoritative guidance and literature highlights the novelty of this topic and that legal and scholarly consideration of the status of synthetic data is at an early stage in the UK, Europe and around the world.

In terms of case law and precedent, we have been unable to find any court judgment or opinion addressing the question of whether synthetic data are personal data (or in what circumstances this is the case). This is not surprising given how recently the field of synthetic data generation has developed from research into application. However, the lack of judicial consideration does not mean that authoritative regulatory bodies are yet to consider the topic completely. Searching of the activities and outputs from data protection authorities across Europe reveals that synthetic data are very much on the radar of data regulators.

Current Data Protection authorities' approaches to synthetic data

In our review of activities and outputs from data protection authorities and other authoritative bodies across Europe, we identify significant focus on synthetic data all within the last three years.

CNIL (France)

A series of articles were published by the French data protection authority, the CNIL, focused on synthetic data in October 2022. The author, expert engineer Alexis Léautier, discusses the potential of synthetic data generation in terms of the use cases that arise including training of AI models, software testing, data sharing and the benefits, such as improved confidentiality, completeness, specificity or accuracy, that synthetic data may bring to these uses.[75]

The origin and development of different synthetic data generation techniques is also considered in detail as well as the potential combination with additional privacy techniques such as differential privacy to provide formal guarantees against re-identification. However, in a second article, there is a much stronger focus on the potential privacy or identification risks that can arise from synthetic data generation. These include risks of membership inference or attribute inference on sets of synthetic data generated by several algorithms (sampling, Bayesian network, GAN) from census and hospitalization data.[76]

The CNIL article also considers evidence of the possibility of membership inference attacks on GANs reported by Chen et al.[77] but also emerging research on methods that can be used to guard against such attacks or apply further privacy guarantees while retaining high data utility (e.g. differential privacy).

Overall, these articles suggest a cautious welcome from the French regulator, acknowledging that risks associated with the use of synthetic data are not zero, but also that their use is less risky than the use of real personal data.

Daten Ethik Komission (Germany)

Although not a data protection regulator, the statutory German Federal Data Ethics Commission was tasked with consideration of appropriate ethical and legal standards for the use of data in algorithmic systems (including AI) in 2018. In its comprehensive final report presented to the Federal Government of Germany on October 23 2019, the Commission found that research in the field of synthetic data showed enormous promise, and recommended that more funding should be provided for this area.[78]

---

75 Léautier A. [Données synthétiques] - Dis papa, comment on fait les données? Laborotoire d'Innovation Numerique de la CNIL, 18 October 2022. Available at: <https://linc.cnil.fr/fr/donnees-synthetiques-dis-papa-comment-fait-les-donnees-12> accessed 27 February 2023

76 Stadler T, Oprisanu B, Troncoso C. Synthetic data–anonymisation groundhog day. In31st USENIX Security Symposium (USENIX Security 22) 2022 (pp. 1451-1468)

77 Chen D, Yu N, Zhang Y, & Fritz M. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. Proceedings of the ACM Conference on Computer and Communications Security. 2020 343–362. Available at: <https://doi.org/10.1145/3372297.3417238> accessed 27 February 2023

78 German Federal Data Ethics Commission. Opinion of the Data Ethics Commission. 23rd October 2019. Available at: <https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/datenethikkommission-abschlussgutachten-lang.pdf;jsessionid=C0CC81C4893FC8EBC3F9606D3B1CBF59.2_cid364?__blob=publicationFile&v=5> accessed 27 February 2023

On one hand the Commission asserted that if a set of synthetic data contains no references to persons, it is anonymous and does not fall within the scope of the GDPR. However, the Commission also highlighted a lack of legal certainty in a number of different areas, for example concerning the anonymisation and pseudonymisation of data, the identification and consideration of a link between individuals and (allegedly anonymised) data sets, and synthetic data.

Datatilsynet (Norway)

The Norwegian data protection authority has been widely reported as endorsing the use of synthetic data when testing new technologies and software (clearly recommended or even mandated in their guidance)[79] over real personal data.[80]

This reporting is partly due to an enforcement decision by the Authority against the Norwegian Confederation of Sport (NIF) resulting in a EUR 125,000 fine for making the personal data of 3.2 million Norwegians available online for 87 days as a result of an error in connection with testing of a cloud computing solution. In the decision, the Data Protection Authority found that the testing could have been achieved by processing synthetic data instead and the DPA strongly recommended the use of fictitious data to mitigate the risk 'considerably'.

The implication of this decision is that use of appropriate synthetic data would not have involved disclosure of personal data in the case at hand. The scale of the fine suggests a large incentive for those testing software and other tools to adopt synthetic data to mitigate or completely eliminate privacy risks.

European Data Protection Supervisor (EDPS)

The European Union's own internal data protection authority, the European Data Protection Supervisor (not to be confused with the European Data Protection Board which provides guidance to all DPAs across the EU) is responsible for assessing the compliance of EU bodies and organisations with the GDPR. Although this makes it just one of many peer DPAs across Europe, the close relationship between the EDPS and EU institutions and the pan-EU scale of the activities it governs tends to elevate the authority of the EDPS outputs.

In 2021 the EDPS foresight initiative TechSonar selected Synthetic Data as one of the key technologies for consideration. In the related Report and associate blog, the EPDS identify positive foreseeable impacts of Synthetic Data, including added value for privacy of individuals 'whose personal data does not have to be disclosed' as well as potential negative impacts'.[81] These include risks of identification where synthetic

---

79 Datatilsynet. Programvareutvikling med innebygd personvern. Available at: <https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/innebygd-personvern/programvareutvikling-med-innebygd-personvern/test/> accessed 27 February 2023

80 Deloitte. Syntetiske testdata eller personopplysninger ved testing? 24th August 2020. Available at: <https://www2.deloitte.com/no/no/pages/legal/articles/syntetiske-testdata-eller-personopplysninger-i-test.html> accessed 27 February 2023

81 European Data Protection Supervisor (EDPS). TechSonar 2021-2022 Report. Doi:10.2804/641632; EDPS Websitse. Synthetic Data. Available at: <https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en> accessed 27 February 2023

datasets closely mimic real data, risk of membership inference attacks and a lack of clarity about other privacy risks from generative models in particular at this early stage of development.[82] In terms of current practice, the EDPS article advises that:

> 'A privacy assurance assessment should be performed to ensure that the resulting synthetic data is not actual personal data. This privacy assurance evaluates the extent to which data subjects can be identified in the synthetic data and how much new data about those data subjects would be revealed upon successful identification.'

Perhaps the most important consideration of Synthetic Data led by the EDPS was an Internet Privacy Engineering Network (IPEN) Webinar on the 16 June 2021 entitled: "Synthetic data: what use cases as a privacy enhancing technology?".[83] The workshop focused on the use of "synthetic data" as a possible technology to mitigate data protection risks. We discuss some of the contributions to this webinar further below, following consideration of the UK DPA's approach to Synthetic Data so far.

The Information Commissioner's Office (UK Data Protection Authority)

The UK data protection authority, the Information Commissioner's Office or ICO, is a leading authority for horizon scanning and consideration of emerging technologies.[84] It is not surprising therefore that the ICO has touched on synthetic data generation in several of its activities. In version 1.0 of its guidance on using AI and personal data appropriately and lawfully, synthetic data is mentioned as a means of preserving privacy while developing AI systems[85] and the latest updated guidance on AI and Data Protection (updated 15th March 2023) refers to synthetic data as a privacy enhancing technology (PET). In this guidance the ICO makes three main points about synthetic data.[86]

First, that the development of models using synthetic data can help to preserve privacy and '[t]o the extent that synthetic data cannot be related to identified or identifiable living individuals, it is not personal data and therefore data protection obligations do not apply when you process it'. Second, AI developers are reminded that the generation of synthetic data will generally involve processing some real data and where this can

82 EDPS. TechSonar 2021-2022 Report. p11

83 EDPS Website. IPEN Webinar 2021 - "Synthetic data: what use cases as a privacy enhancing technology?" Available at: <https://edps.europa.eu/data-protection/our-work/ipen/ipen-webinar-2021-synthetic-data-what-use-cases-privacy-enhancing_en> accessed 27 February 2023

84 ICO. Research and Reports- Tech Horizons Report. Available at: https://ico.org.uk/about-the-ico/research-and-reports/tech-horizons-report/> accessed 27 February 2023

85 ICO. How to use AI and personal data appropriately and lawfully. 20221108 Version 1.0. Available at: <https://ico.org.uk/media/for-organisations/documents/4022261/how-to-use-ai-and-personal-data.pdf> accessed 27 February 2023

86 ICO. Guide to Data Protection: How should we assess security and data minimisation in AI? Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/#whatstepsshould> accessed 27 February 2023

be related to identifiable individuals the processing of such data must comply with data protection laws. A third point addresses the residual privacy risks that may remain. In some cases, the ICO suggests, 'it may be possible to infer information about the real data which was used to estimate those realistic parameters, by analysing the synthetic data.' The example given is of outlier individuals who can be identified in both the training and synthetic data. A potential trade-off between utility and privacy is acknowledged with the warning that 'avoiding such re-identification may require you to change your synthetic data to the extent that it would be too unrealistic to be useful for machine learning purposes'.

Other aspects of this very recent guidance are relevant to the assessment of whether, or in what circumstances, synthetic data may be personal data, such as the discussion of relevant forms of attack that could be made on AI models and their outputs. We will consider these more fully below.

Perhaps the most important consideration given to synthetic data by the ICO is in draft anonymisation guidance published in September 2022 on Privacy-enhancing technologies (PETs).[87] In this draft guidance, synthetic data are considered alongside other PETs such as differential privacy or homomorphic encryption. The ICO define synthetic data as:

> 'artificial' data generated by data synthesis algorithms, which replicate patterns and the statistical properties of real data (which may be personal data).'

A distinction is made between partially synthetic data (which only synthesises some variables of the original data) and fully synthetic data where all variables are synthesised, although the ICO does not explicitly discuss the different privacy implications of these two forms of data. However, a number of privacy advantages of synthetic data are discussed. For example, synthetic data may allow large datasets to be created from small real datasets and therefore help to promote the principle of data minimisation.

Perhaps most importantly, it is implied that synthetic data generation can be used to generate 'non-personal data', and that the ICO would not consider all synthetic data to remain or constitute 'personal data':

> 'You should consider synthetic data for generating non-personal data in situations where you do not need to, or cannot, share personal data.'

87 ICO. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Chapter 5: Privacy-enhancing technologies (PETs). September 2022

However, this does not amount to an endorsement that synthetic data are not 'personal data' and the ICO acknowledges that this is an active research area and that there are privacy risks associated with the use of synthetic data. These include a trade-off between greater utility/mimicking of real data and a greater likelihood of revealing individuals' personal data, and the potential for attacks such as 'model inversion attacks'. The ICO's guidance is that:

> 'You should consider whether the synthetic data you generate is personal data. You should focus on the extent to which individuals are identified or identifiable in the synthetic data, and what information about them would be revealed if identification is successful.'

The difficult balance between utility and privacy is recognised by the ICO in terms of additional safeguards that could be applied in combination with synthetic data (for example to protect against singling out). In particular, that '[u]sing differential privacy with synthetic data can protect any outlier records from linkage attacks with other data. However, it may reduce the utility of the data and introduce a degree of unpredictability regarding the preservation of data characteristics.'

Summary of DPA approaches

Overall, the approach of the data protection authorities to synthetic data generation have been consistent in one regard: They have all approached synthetic data as a potential Privacy Enhancing Technology (PET). This is perhaps unsurprising given the role and function of these authorities, but it is also potentially significant that not all synthetic data generation is driven by a desire to enable data processing or sharing which would otherwise only be feasible using personal or private information. It could be argued that viewing synthetic data generation purely through the lens of Privacy Enhancing Technologies places a greater emphasis on their privacy risks as well as strengths than might be appropriate in all cases. We consider this issue more fully below, following scrutiny of how legal scholars are approaching synthetic data within data protection law.

Legal scholarship and commentary on synthetic data

Our searching of legal literature has found only very limited consideration thus far of the question of the status of synthetic data in data protection law. There has been considerable focus on the privacy implications and threats to various modes of synthetic data generation in the technical literature[88] with frequent reference to legal standards in those articles.

88 E.g. Chen D, Yu N, Zhang Y, & Fritz M. (2020). GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. Proceedings of the ACM Conference on Computer and Communications Security, 343–362. https://doi.org/10.1145/3372297.3417238; Stadler T, Oprisanu B, Troncoso C. Synthetic data–anonymisation groundhog day. In31st USENIX Security Symposium (USENIX Security 22) 2022 (pp. 1451-1468); Yoon, J., Drumright, L. N., & Van Der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). IEEE Journal of Biomedical and Health Informatics, 24(8), 2378–2388. Available at: <https://doi.org/10.1109/JBHI.2020.2980262> accessed 27 February 2023

There has been some consideration of how synthetic data may fit within U.S. privacy laws, from Bellovin and colleagues in 2019.[89] However, there has been no direct consideration of the question of when, or in what circumstances, synthetic data may constitute personal data from a legal perspective in Europe or the UK. This not to say that there are no legal articles that refer to synthetic data. However, these are currently tangential references to synthetic data as part of a range of wider legal debates including those on the appropriate interpretation of the scope of 'personal data' in EU data protection law, the appropriate role and function of data protection law and, appropriate responses to artificial intelligence and automated processing within data protection law and the regulatory framework more widely.

Although the literature is yet to fully consider this legal question, there have been recent contributions to the debate from a legal perspective in workshops and conferences. Notably, the EDPS IPEN webinar held in 2021 on Synthetic data involved discussion of synthetic data and the application of EU data protection law from technical and legal perspectives.[90]

Although not speaking directly to interpretation of the law, Professor Khaled el Emam (Ottowa University), a leading authority on anonymisation of health data, PETs and data synthesis, provided an overview of the rapidly developing field of data synthesis. He expressed confidence that it was currently feasible (or would be in the very near future) to test and quantify privacy risks for all the main synthetic data generation methods and that, by combining data transformations and additional security, privacy or contractual controls, the residual privacy risk for fully synthetic data could be reduced to well under established thresholds in the context of biomedical data sharing (e.g., well below 0.09 probability).

Perhaps the most relevant query raised by Prof. el Emam is the extent to which an exceptional approach might (without justification) be adopted for synthetic data and relevant models. He contested that commonplace reporting of regression models in academic literature gives rise to similar forms of privacy risk (presumably through forms of model inversion attack as discussed by Fredrikson and colleagues for example)[91] and queried why synthetic data should be treated differently and presumptively not made public.

Consideration: There should be consideration of the extent to which standards applied to privacy of synthetic data are consistent with other forms of data generation such as statistical modelling.

---

89 Bellovin SM, Dutta PK, Reitinger N. Privacy and synthetic datasets. Stan. Tech. L. Rev.. 2019;22:1

90 EDPS Website. IPEN Webinar 2021 - 'Synthetic data: what use cases as a privacy enhancing technology?' Available at: <https://edps.europa.eu/data-protection/our-work/ipen/ipen-webinar-2021-synthetic-data-what-use-cases-privacy-enhancing_en> accessed 27 February 2023

91 Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In 23rd {USENIX} Security Symposium ({USENIX} Security 14, 2014 (pp. 17-32)

A further session at the EDPS organised Webinar on 'synthetic data as a privacy enhancing technology' involved insights from a more legal perspective. Speaking from a legal data protection perspective Dr Dara Hallinan, a specialist in data protection law and health research, and Kelsey Finch, former senior Counsel at the Future of Privacy Forum, provided an early indication of the considerations that we might expect to see emerging in legal literature on this topic in the near future. Both highlighted the way in which synthetic data could push at the boundaries of data protection law in terms of the form of harms that it should address and the nature of threats that the legal framework is intended to guard against.

Finch emphasised the importance of tracking real world uses and consequences of synthetic data and therefore adopting a relatively precautionary approach to this new technology. However, she also emphasised that regulatory responses should not view all synthetic data in the same way and that fully synthetic datasets without 1-1 matches with the training data might, for example, pose different risks to hybrid real-synthetic datasets.

As Prof. el Emam also discussed, Finch emphasised that context and practice will have a major influence on the risk of re-identification that may arise, and that policymakers should not assume that all forms of data synthesis are equally effective; it is critical that policymakers define the parameters that influence the level of risk that arises from synthetic data generation and consider domain specific guidance to achieve the best balance between utility and privacy in context.

Dr Hallinan posed some fundamentally challenging questions for the application of data protection law to synthetic data generation. Dr Hallinan suggested that synthetic data may display characteristics which could make them very different to other anonymisation techniques and argued that another lens could perhaps be appropriate.

In his opinion, viewing synthetic data solely from an anonymisation or privacy enhancing technology perspective could lead to misleading or incomplete conclusions from a regulatory or policy perspective. There may be sufficiently novel and different aspects to this form of data creation compared with established anonymisation techniques which give rise to an unclear relationship between the synthetic data and original records and potentially different risks and challenges.

For example, Hallinan highlighted, it would be very unclear how data protection principles such as the principle of accuracy should be considered in relation to synthetic data if it is considered personal data. Alternatively, some potential harms would be inadequately dealt with through data protection regulation, such as the potential subversion of the Nagoya Protocol on benefit sharing from use of genomic resources if scientists are able to create purely synthetic genomes instead of obtaining the permission of specific (often indigenous) populations. Equally, the generation of synthetic group data profiles would also potentially fall outside the individualistic framework of data protection law.

Consideration: regulators and authorities should carefully consider whether data synthesis presents a novel form of data creation and processing compared with established anonymisation techniques, and the extent to which it gives rise to novel risks and challenges.

Finally, a 2022 workshop paper authored by Cesar Augusto Fontanillo López and Abdullah Elbi from the Center for IT and IP Law KU Leuven considers the legal nature of synthetic data in the European context.[92] In this thorough consideration of the legal qualification of synthetic data under the GDPR, the authors highlight considerable uncertainty in both the interpretation of relevant legal concepts, such as identifiability, and their application to the range of synthetic data generation models that arise in context. These authors note that 'that the bar of anonymisation has been set very high by the European legislator' and conclude that a range of possibilities arise, with synthetic data capable of being considered personal, 'pseudonymous' or anonymous depending on interpretation and context.

92 Fontanillo López CA, Elbi A. On the legal nature of synthetic data. InNeurIPS2022, Location: New Orleans 2022 Nov 3

Summary

It is too early to say with certainty how courts and tribunals in the EU or UK will approach the question of whether, or in what circumstances synthetic data may qualify as personal data. There are yet to be any judgments or decisions on this issue. However, some of the data protection authorities in Europe and the UK Information Commissioner's Office have set out preliminary considerations for synthetic data generation, primarily as part of the assessment of emerging privacy enhancing technologies and AI processing.

While all these bodies acknowledge the potential privacy benefits of synthetic data over 'real' personal data processing, all the relevant outputs and guidance acknowledge—and highlight, to differing degrees—potential privacy threats to synthetic datasets and the models used to generate them. In particular, the authorities identify a relationship between the utility of synthetic data and privacy risks and specific forms of threat, such as membership inference attacks and model inversion attacks. The ICO and other authorities consequently recommend additional privacy enhancing methods such as differential privacy, where feasible to safeguard synthetic data and relevant models when released.

The combination of considering synthetic data primarily as a privacy enhancing technology and current uncertainty about the threat level posed by forms of attack that may be used against synthetic datasets and models, results in a generally precautionary approach thus far from regulators.

What is therefore clear, is that the data protection authorities are not approaching synthetic data as presumptively 'non personal data'. Instead they are adopting what could be termed an 'orthodox' approach to synthetic data as a novel (privacy enhancing) technology which begins with the position that data being processed are 'personal data' and will only reach a different conclusion with a high degree of confidence that threats of re-identification are minimal and well safeguarded.

Given this approach, a crucial question for data controllers is how synthetic datasets and synthetic data generating models may be sufficiently protected against risks of identification, that it can be concluded with confidence that 'personal data' are no longer being processed. Unfortunately, answering this question in relation to any form of data is not straightforward.

Authorities and courts across Europe have traditionally adopted different standards and approaches to identifiability and anonymisation (in practice if not in formal terms) and this variation has prevailed even under the 'harmonising' force of the GDPR. To some degree this uncertainty is alleviated in the UK where the ICO and domestic courts' approach are formally the only relevant considerations but there are a range of practical and policy factors (as we discussed in Section 3) which suggest that the UK approach to identification and anonymisation will not likely diverge heavily from those in other major jurisdictions, including the EU in the near future.

This question of the appropriate approach to identifiability, anonymisation and the scope of 'personal data' is a major topic of legal academic debate.[93] Although there is very limited published literature considering synthetic data within this framework in the European context, there are some scholarly contributions in blogs and workshops papers which follow the 'orthodox' approach to determining whether synthetic data are personal data. These commentators highlight the extent to which answers are contingent on contested aspects of data protection law, including fundamental questions about the appropriate scope and role of this part of the regulatory framework.

However, as Dara Hallinan argues there may be reasons to believe that a more novel approach would be appropriate to the governance of synthetic data models and outputs, in certain circumstances. These partly relate to the nature of synthetic data as distinct from source data in potentially wholly novel ways when compared with other 'anonymisation' technologies and also wider questions about the nature of the harms that could arise from synthetic data processing and whether, or how well, those harms should be addressed through data protection law as opposed to other parts of a regulatory framework. These questions form part of much wider debates about the nature of privacy and how novel AI-driven technologies should be governed and regulated. We will consider them further in Section 5 but first we consider how the 'orthodox' approach to 'personal data' is likely to apply to synthetic health data generation.

93 Dalla Corte L. Scoping personal data: towards a nuanced interpretation of the material scope of EU data protecton law. European Journal of Law and Technology. 2019 May 16;10(1); Purtova N. From knowing by name to targeting: the meaning of identification under the GDPR. International Data Privacy Law. 2022 Aug;12(3):163-83.; Finck M, Pallas F. They who must not be identified—distinguishing personal from non-personal data under the GDPR. International Data Privacy Law. 2020 Feb 1;10(1):11-36

What is the likely regulatory approach?

- The current ‘orthodox’ approach
- An alternative approach?
- Summary

5. What is the likely regulatory approach?

In light of our findings in the previous section in relation to the current approach of data protection authorities and legal scholars to synthetic data, we can identify a set of factors that will almost certainly be key to determining whether, or in what circumstances synthetic data will be considered personal data in the short-to-medium term.

The current 'orthodox' approach

As we identified in the previous section, a current 'orthodox' approach to assessing the status of synthetic data is discernible. This is a cautious approach which begins with the position that if data being processed are 'personal data', a different conclusion will only be reached with a high degree of confidence if it can be demonstrated that threats of re-identification are minimal and well safeguarded (i.e. that data have been effectively anonymised). The key initial question is therefore whether personal data are being processed from the outset.

Are the source or training data 'personal data'?

The first critical factor is whether synthetic data are generated using 'personal data'. If no personal data are involved at the outset then it is highly unlikely that the synthetic dataset will be considered as potential personal data. Strictly speaking, this is not impossible because it is feasible that synthetic data could be generated which purely by chance, matches an individual more closely than the source or training data. Since we do not yet know whether such chance matching would fall outside the scope of data protection law, we cannot say with absolute certainty that synthetic data generated from non-personal data will always be considered non-personal.

However, in terms of the development of synthetic data using real patient datasets (such as in the examples provided in Section 2) a chain of processing clearly begins with identifiable personal data. Matters are slightly complicated in the health data context by the way in which patient data is frequently processed from collection to reduce risks of identification and preserve privacy. For example, by removing direct identifiers (names, addresses and NHS numbers for example) and coding or 'pseudonymising' the data so that an individual can only be readily re-identified by someone who has the key or code to connect their data with their identity.

As we discussed in Section 3, since the GDPR explicitly incorporated 'pseudonymisation' for the first time there has been significant academic debate as to whether such pseudonymised data always constitute 'personal data'.[94] In particular, there have long been different approaches across the EU to the status of pseudonymised data in the hands of a third party who does not have access to the key or code. In some jurisdictions, such data were not considered to fall outside the scope of 'personal data' but in the UK and the Netherlands, for example, the regulators viewed such data as capable of falling outside the scope of personal data in certain circumstances.[95]

The UK's departure from the European Union has provided greater scope for, and likelihood of, a clear and consistent approach from the courts and regulators that pseudonymised data are capable of becoming anonymous, in third party hands. This is explicitly endorsed in the latest draft guidance from the ICO.[96] However, as the ICO cautions in this draft guidance, it cannot be assumed that pseudonymous data become anonymous in the hands of a third party (i.e. the party developing synthetic data). Several things need to be taken into account to assess the likelihood of identifiability, considering things like the cost of and time required for identification, and the state of technology at the time of the processing, and the techniques and controls placed around the data once in the recipient's hands.[97]

Consideration: Unlike some EU jurisdictions, the ICO and courts' approach to pseudonymisation provides room for pseudonymised data to be considered anonymised or outside the scope of 'personal data' where separated from a key and in another party's hands.

It is outside the scope of this report to consider the status of patient data that has or can be used to generate synthetic data and we will continue our analysis on the assumption that such data are considered personal (albeit likely pseudonymised) data.

---

94 Mourby M, Mackey E, Elliot M, Gowans H, Wallace SE, Bell J, Smith H, Aidinlis S, & Kaye J. (2018). Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. Computer Law and Security Review, 34(2), 222–233. Available at <https://doi.org/10.1016/j.clsr.2018.01.002> accessed 27 February 2023

95 Shabani M. Borry P. Rules for processing genetic data for research purposes in view of the new EU General Data Protection Regulation review-article. European Journal of Human Genetics. 2018. 26(2), 149–156

96 ICO. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Chapter 3: Pseudonymisation. February 2022. Available at: <https://ico.org.uk/media/about-the-ico/consultations/4019579/chapter-3-anonymisation-guidance.pdf> accessed 27 February 2023

97 Ibid, p5

Assessing the identifiability of output synthetic data and synthetic data models

If the chain of processing begins with identifiable and personal data, this is likely to establish a presumption that data remain personal data, unless effective anonymisation can be shown to have taken place. As we have discussed in earlier sections, determining whether data are identifying in data protection law terms involves assessment of the means that are reasonably likely to be used to identify an individual, the data and its environment, the context, scope and purposes of the processing; and technical and organisational measures applied to safeguard the data.[98]

For those developing and using synthetic health data in the UK, the ICO and other relevant authorities (such as the NHS Health Research Authority) arguably adopt a more achievable standard than some counterpart authorities in Europe in terms of anonymisation. Crucially, the ICO (and in the past the courts[99]) have suggested a more pragmatic approach than the Article 29 Working Party Opinion 05/2014, which suggested that de-identification would need to be irreversible to achieve anonymisation.

We are waiting for the ICO and the European data protection body, the EDPB, to update their guidance on anonymisation and in the ICO's case, the hugely influential Code of Practice on Anonymisation which was published in 2012.[100] Although organisations are still using this Code to guide their assessment of anonymisation, it is now significantly out of date in terms of both the underlying legal framework and the tools and technologies, including synthetic data, that are available.

Fortunately, as we have already noted, the ICO has produced draft guidance on anonymisation which, although subject to change, strongly indicates the approach that the ICO would like to adopt.[101] A final version of this and separate PET guidance is impending and there are very promising indications for researchers in the ICO's intent to produce a new Chapter on anonymisation in research and to publish several case studies to demonstrate how various technologies can be used to facilitate compliant data sharing.

---

98 ICO. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Chapter 2: How do we ensure anonymisation is effective? October 2021. Available at: <https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf> accessed 27 February 2023

99 *Department of Health v Information Commissioner* [2011] EWHC 1430 (Admin) (Department of Health)

100 ICO. Anonymisation: managing data protection risk code of practice. 2012. Available at: <https://ico.org.uk/media/1061/anonymisation-code.pdf> accessed 27 February 2023

101 ICO. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Available at: <https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-call-for-views-anonymisation-pseudonymisation-and-privacy-enhancing-technologies-guidance/> accessed 27 February 2023

The ICO's draft guidance strongly endorses a risk-based approach to identifiability and anonymisation under the UK GDPR. This distinguishes between 'truly anonymous' data which is impossible to be used to re-identify an individual and 'effectively anonymised' data where 'identifiability risk is sufficiently remote.'[102] The standard required for data to fall outside the scope of 'personal data' is only for data to be considered effectively anonymised. What is required is a contextual risk assessment of the place of synthetic data on a 'spectrum of identifiability'. If data are considered to be unlikely to be identifiable because the risk of identification in context is sufficiently remote, the information can be considered effectively anonymised. There are a number of generic factors that can be considered alongside more specific aspects for assessment of synthetic data and risk of identification.

General factors: Singling out, inferences and linkability

As we note in Section 3, three forms of identification risk are particularly prominent in data protection guidance. Singling out refers to the possibility of identifying the records of an individual separate from others within a dataset.[103] This risk could be lower for synthetic data than other forms of data because synthetic data generation does not just seek to remove identifiers from a dataset but it also results in changes or differences in the data points in a record compared with the source dataset. However, synthetic patient data of the kind we consider in Section 2 does incorporate 'individual' level data so there are potential risks of singling out a real individual if the data can be related to a living individual in ways that we discuss below.

Linkability is another risk that should be assessed and it relates to the potential combination of otherwise innocuous information with additional available data to lead to the identification of an individual. Finally, inferences may allow the identification of an individual. This refers to the potential to infer or predict using deductive logic, that an individual is part of a dataset based on querying of a model, comparison with other pieces of information or specific knowledge about the individual (among other things).

In the context of synthetic patient data, the primary risks of identification relate to inferences that may be made from observing the data, the model and other sources of information (for example the ground truth data if that is available) but it is also feasible that some forms of synthetic data generation give rise to singling out and linkage risks as well (e.g. if data are only partially synthetic) especially for outliers.

---

102 ICO. Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance. Chapter 2: How do we ensure anonymisation is effective? October 2021. p9

103 The risk of identification singling out creates has generally been sufficient for authorities and courts to consider such data as 'personal data'. However, we have argued in previous work that something further is always required to connect that data with a living individual by content, purpose or effect. Mitchell C, Ordish J, Johnson E, Brigden T, Hall A. The GDPR and Genomic Data: PHG Foundation report on the impact of the General Data Protection Regulation and Data Protection Act 2018 on Regulating Genomic Technologies in Healthcare for the Information Commissioner's Office. PHG Foundation. 2020

Specific factors: Risks that may arise in the context of synthetic data generation

Based on the literature and guidance, some specific privacy or identification risks may arise in the context of synthetic data generation:

- where data are only partially synthetic by design, singling out, linkage with other sources or inferences may be used to identify an individual
- depending on the method, it is feasible that no change or insignificant change has been made to an individual record (i.e. only minor jittering) so that a record is left insufficiently altered
- alternatively, the generation method could inadvertently incorporate the real (ground truth or training) data into the output dataset. This can be referred to as 'leakage'.
- a model could be vulnerable to 'overfitting', where it pays too much attention to the details of the training data, 'effectively remembering particular examples from the training data'. This can occur where there are too few examples in the training data for example.[104] This can be exploited by model inversion and membership inference attacks.[105]
- a synthetic data model could be vulnerable to 'model inversion attacks', where attackers already have access to some personal data belonging to specific individuals in the training data, but can also infer further personal information about those same individuals by observing the inputs and outputs of the machine learning model.[106]
- a synthetic data model may be vulnerable to membership inference attack, which allows actors to deduce whether a given individual was present in the training data of a machine learning model.[107]

In terms of the risks that may apply to the synthetic data model, a distinction can be made between 'black box' and 'white box' attacks. In the former case, the attacker only has the ability to query a model and observe relationships between inputs and outputs. In a white box attack, the adversary also has full access to the model's structure and

104 ICO. Guide to Data Protection: How should we assess security and data minimisation in AI? Available at: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/#whatstepsshould> accessed 27 February 2023

105 Ibid

106 ICO. Glossary. Model inversion attack. Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/glossary/> accessed 27 February 2023

107 ICO. Glossary. Membership inference attack. Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/glossary/> accessed 27 February 2023

parameters.[108] Although preventing access to the model itself could limit some risks, it has been shown that both membership inference and model inversion attacks can be achieved for some forms of machine learning models through a black box attack only.[109]

Consideration: Data controllers should be aware that, depending on method, there are several potential privacy or identification risks that may arise in the context of synthetic data generation some of which may not require full access to a model to achieve.

General factors: The data environment and contextual controls

While there are a range of risks and threats that could apply to synthetic health data and synthetic data models, depending on the methods involved and the nature of the data, a range of important factors relate to the wider technical and organisational (and legal) measures in place around the data. This is termed the 'data environment' and the ICO's draft anonymisation guidance expressly refers to a range of factors involved including:

- additional data that may exist (e.g. other databases, personal knowledge, publicly available sources); who is involved in the processing, and how they interact;
- the governance processes that are in place to control how the information is managed (e.g. who has access to it and for what purposes); and
- the legal considerations that may apply, such as:
  - any gateways that may impact the potential for disclosing information that enables individuals to be identifiable; or
  - prohibitions that mean while information could technically be combined to aid identifiability, doing so is against the law (e.g. professional confidentiality).

Specific factors: the data environment for synthetic health data

Due to the sensitivity of the data involved, health data is a highly regulated and carefully governed domain. In the UK, a range of safeguards are routinely deployed to protect patient data within and beyond the NHS, including pseudonymisation, disclosure control, professional ethical codes, duties of confidentiality and prohibitions on data sharing without lawful basis. Where such data are used in the context of health research, a range of additional safeguards are highly likely to apply including data use agreements (or equivalent legal agreements between data custodians and users) and technical safeguards, including secure research environments where data cannot be brought in or out.

---

108 Veale M, Binns R, Edwards L. Algorithms that remember: model inversion attacks and data protection law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018 Nov 28;376(2133):20180083

109 Ibid, pp5-6

In this context, the generation of synthetic data and use of output synthetic datasets could be subject to a spectrum of these measures and safeguards according to the risk assessment of the controller.

Consideration: Data controllers will need to carefully consider the technical nature of the data they generate as well as the environment of legal and organisational controls surrounding the data as part of any risk assessment.

General Factors: New threats, technologies and other developments

Finally, a key element in assessing and mitigating identifiability relates to the dynamic nature of 'personal data'. As discussed in relation to the definition of 'personal data' in the UK GDPR, data that at one point is effectively anonymised can, in time, become personal data as new threats and technologies emerge. This may suggest a precautionary approach, particularly when applied to synthetic data generation and the current uncertainty about the nature and real significance of the threats posed.

Consideration: the open release of even fully synthetic health data may still present an unforeseen privacy and identification risk. Some technical and organisational safeguards may be required to prevent these arising.

Determining when the risk threshold of identifiability for 'personal data' is met?

While a range of general and more specific factors can be identified, the difficulty for any contextual assessment of identifiability within data protection law is that there are no legally binding, agreed or quantified standards against which threats can be assessed. There are some objective factors set out by the ICO. The latest draft guidance states that:

> 'You should approach assessing identifiability risk by considering what is reasonably likely relative to the context. This includes whether identification is technically and legally possible, taking into account objective criteria including:
>
> - how costly identification is in human and economic terms;
> - the time required for identification; and
> - the state of technological development at the time of processing (i.e. the techniques you use for anonymising the data, and/or when you are sharing the dataset with another party); and
> - future technological developments (i.e. as technology changes over time).'[110]

110 ICO Draft anonymisation guidance, Chapter 2, p12-13

However, it is ultimately a matter of judgement for the data controller to determine whether they believe it is no longer reasonably likely that data will be capable of identifying an individual. The ICO's draft guidance refers to a risk of identifiability that is sufficiently remote for data to be considered effectively anonymised but neither ICO guidance nor case law can set a clear standard in practice for all circumstances.

Nevertheless, there are sources of guidance and best practice standards for de-identification and anonymisation which are designed to provide practical assistance towards achieving legal compliance. A notable example is the UK Anonymisation Network's (UKAN) Anonymisation Decision Making Framework (ADF).[111] This represents a comprehensive effort to bridge the gap between legal ambiguity and practical assessment and it provides structured guidance for data controllers in both assessing and implementing privacy and security measures and environments. (Some of these aspects also overlap the more general guidance for a data protection impact assessment or DPIA. For further practical guidance in relation to DPIA's and the ADF see the Annex to this report).

The UKAN's approach incorporates all of the general factors set out above and builds on them with applied guidance as part of what it terms a data situation audit. The 'data situation' is an 'aggregate set of relationships between some data and the set of their environments. It is a combination of the factors that relate to the data itself and environmental factors that, as the ADF reminds users, is frequently dynamic as data moves between actors and into other environments. The ADF sets out an iterative process that data controllers can follow to audit their data situation and determine whether to employ disclosure risk assessment and proportionate controls. The risk threshold used by the UKAN is that risk should be 'negligible' for effective anonymisation and the ADF strongly emphasises that anonymisation is a continuing process requiring ongoing assessment and adjustment over time.

While the ADF provides a practical framework for considering and assessing the identifiability of data, a range of options for assessment will need to be used by data controllers to scrutinise their data and its environment. Most prominent among these is penetration or intruder testing. The ICO refers to a 'motivated intruder' test.

The ICO recommends that a data controller adopts a 'motivated intruder' test as part of their risk assessment. It is a test that considers whether an ordinary person or a determined person with a particular reason to want to identify individuals, used by both the ICO and the Information Tribunal, which hears DPA 2018 and FOIA appeals.

---

111 Elliot M, Mackey E, O'Hara K, The Anonymisation Decision Making Framework: European Practitioners' Guide. 2nd ed. Manchester: UNAN, University of Manchester; 2020. Available at: <https://ukanon.net/framework/> accessed 27 February 2023

It assumes that a motivated intruder is someone that is reasonably competent, has access to appropriate resources; and uses investigative techniques but the ICO also acknowledges that public release of data and the particular attraction or sensitivity of certain data could heighten the threat.[112]The Office for National Statistics also provides practical guidance on how to undertake motivated intruder testing.[113] In the context of synthetic health data, penetration testing might involve internal testing by skilled team members with access to relevant data, models and additional tools and information.

Consideration: Controllers should follow best practice approaches to assessing privacy risks, such as the UK Anonymisation Network's Decision Making Framework and, where relevant, carry out penetration or motivated intruder testing.

Applying additional technical, organisational or legal safeguards

Finally, the risks of identification that may arise in some forms of synthetic data generation could also be mitigated through the application of additional technical, organisational or legal safeguards. This could shift the risk level sufficiently that data fall outside the scope of 'personal data'. In general, options for such mitigations include controls on the data such as suppression or noise addition, and controls on the environment such as access controls/agreements or licensing. It is also feasible to apply additional privacy or confidentiality measures, such as differential privacy to ensure that the privacy risks of any given output are sufficiently low. However, this may further limit the utility of the data.

In the context of synthetic data generated using real patient data, a range of technical and organisational safeguards are likely to be in place surrounding the processing. What is left for data controllers to consider is their level of access or release (on a spectrum from open to highly controlled access via a secure environment) and the extent to which data are adjusted (for example to ensure the removal of outliers) or technically safeguarded using measures such as differential privacy. The choice of methods should be proportionate to the risk.

Consideration: Controllers developing synthetic data from patient data need to consider the extent to which additional privacy measures or technical safeguards, such as differential privacy, may be necessary in combination with organisational safeguards and restricted access, to ensure identification risk levels remain low.

---

112 ICO Draft anonymisation guidance, Chapter 2, p16-17

113 Office for National Statistics. Guidance on intruder testing. Available at <https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/guidanceonintrudertesting> accessed 27 February 2023

In conclusion, a thorough consideration of a range of factors relating to the data, the data environment and relevant mitigations and safeguards will influence whether synthetic data are viewed as personal data or non-personal data. The risk of identification does not need to reach zero in order for this to be the case. Indeed, a form of residual 'risk' is always likely to remain: matching of a synthetic data record with a real individual purely by chance. The better a synthetic data model performs the more likely this form of chance risk is to occur. It is also a dynamic risk which could arise over time if a real person develops certain conditions or health status that comes to match a record in a synthetic dataset.

Whether this form of accidental overlap between a synthetic record and a real person is something that could give rise to 'personal data' is not straightforward. On one hand it could be argued that such information would clearly be about an individual in terms of its content (see Section 4 for this element of 'personal data') and if the synthetic data record can be matched to a real person (as distinct from a group of people) then this would fall within the scope of personal data.

On the other hand, is this chance overlap the sort of 'harm' that data protection law should seek to address? Is there any difference between this and producing a list of plausible mobile phone numbers which by chance include a real number? We return to this conundrum in the following section as part of a consideration whether an alternative approach to some forms of synthetic data and their generation would be appropriate.

An alternative approach?

It is impossible to say with certainty whether any form of synthetic data would be considered personal data in the abstract. As we have discussed, this is because any assessment turns on many factors relating to the data and the context, which will be highly variable depending on the methods used and the purposes of processing. However, on the basis of our analysis of the limited relevant legal guidance, commentary and approaches adopted to emerging technologies, such an 'orthodox' approach is currently likely on the part of the regulators.

As set out above, this approach will begin with a presumption that data generated from 'personal data' will remain identifiable unless effective anonymisation to a remote or negligible level is demonstrated. We term this the 'orthodox approach' because it is the well-established process for scrutiny of anonymisation techniques and privacy enhancing technologies in data protection law and this is the lens through which the authorities are currently viewing synthetic data. However, adopting this approach to all synthetic data generation is not without cost and challenge.

Challenges with the 'orthodox approach'

First, adopting a precautionary approach to synthetic health data generation could 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 the data and restrict how it may be used. This could limit access to high-fidelity synthetic data for public interest purposes, such as the development of novel medical tools and for forms of health research that otherwise are not feasible.

Second, this approach places a high burden on synthetic data developers (and potentially users, who may also have to make an assessment of privacy risks as data controllers) due to the time and expertise required to fully audit identification risks and make consequent adjustments to the data or the data environment. This is likely to slow the availability of synthetic data for health research and development purposes and to potentially diminish work in this area due to the costs involved. Alternatively, costs may be passed on to data users which could suppress access to synthetic data for health and research purposes.

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?

If synthetic data, or even synthetic data models are viewed as within the scope of personal data they will most likely be considered pseudonymised data as they would no longer incorporate identifiers and be separated from additional information which would enable identification. If those developing synthetic data can demonstrate they are no longer able to identify the data subject, Article 11 UK GDPR could provide some relief.[114] In these circumstances data subject rights will not apply but this is not a complete exemption as a data controller should assist a data subject in providing additional information to enable their identification where they seek to do so. Demonstrating full compliance with Article 11 therefore could be challenging and resource intensive.

Consideration: Regulators and policymakers should be aware that adopting a precautionary approach to synthetic data is not without potential cost, including to health research and the development of medical devices in the public interest.

114 UK GDPR, Article 11. Available at: <https://www.legislation.gov.uk/eur/2016/679/article/11> accessed 27 February 2023

Reasons to adopt an alternative approach?

In addition to these potential practical implications of adopting an orthodox approach to all forms of synthetic data generation, there could be other good reasons to consider an alternative approach to some forms of synthetic data which form part of wider debates in literature and policy. For example, a range of questions can be raised about how well synthetic data generation methods and outputs fit within the current data protection framework:

- do machine learning synthetic data generation models really incorporate personal data, i.e. do they 'remember' training data?[115]
- is it better to see some forms of fully synthetic data generation (with controls on overfitting etc.) as a break in a 'chain' of processing that begins with personal data; learning from that data and then throwing it away?
- should purely 'coincidental' matching between synthetic data and a real person be considered a form of identifying inference?

These are open questions which require further consideration by technical experts, regulators and legal scholars. In terms of whether purely 'coincidental' matching should be sufficient to give rise to personal data, there are wider relevant questions about the appropriate role of data protection law within the broader regulatory framework that applies to synthetic health data generation (in particular, those which use AI methodology):

- what harms are actually foreseeable? Coincidental matching? Is that appropriately governed by data protection law?
- are some harms either not governed by data protection law or better addressed by other parts of the legal framework (for example group profiling and predictions)?
- would governing all forms of synthetic health data generation as 'personal data' overstretch the concept, and warp the function of data protection law?

These questions are particularly relevant to AI-driven methods for generating synthetic data where laws and policies are only beginning to be developed to address AI processing and potential harms beyond privacy and data protection concerns. The EU is currently debating a proposal for an AI regulation (the AI Act)[116] which sets rules and limits for a range of technologies, including prohibitions on high risk AI processing.

---

115 Veale M, Binns R, Edwards L. Algorithms that remember: model inversion attacks and data protection law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018 Nov 28;376(2133):20180083

116 European Commission. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. COM/2021/206 final Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206> accessed 27 February 2023

The UK Government, by contrast, has set out its vision for a lighter touch regulatory approach that addresses the use of AI rather than the technology itself, largely through existing sectoral regulations.[117] Put simply, it could be argued that because data protection law is currently the only legal option for regulating many forms of novel and highly sophisticated forms of algorithmic processing, its scope is inevitably being (over)expanded. This could (and perhaps should) change as soon as new parts of the regulatory and governance framework are put in place.

This also highlights the tension at play in the expansive approach to the scope of 'personal data': If almost all information that could be related to an individual comes within the scope of data protection law, actual compliance with the law becomes increasingly difficult, uncertain and challenging.

Default treatment of synthetic data models or fully synthetic datasets with realistically minimal identification risks or even purely chance-based 'risks' of reproducing a real person's attributes in a synthetic dataset could provide examples of such an over-expansion of the scope of 'personal data' and ambit of data protection law. Indeed, in her seminal article 'The law of everything. Broad concept of personal data and future of EU data protection law',[118] Nadezhda Purtova argues that the material scope of the GDPR is growing so broad as to encompass almost anything, including information about the weather. There are counter arguments that this is an over-statement of the current situation[119] and the United Kingdom is potentially on a diverging path from this expansive approach, but the central point remains: Are there better ways to govern certain forms of processing and novel technologies than by stretching the concept of personal data?

Some scholars are already arguing that a different approach is feasible. For example, Bernier & Knoppers argue that separate legislation and regulation should be used to address data-related practices that do not relate to demonstrably identifiable data (such as algorithmic profiling) and that regulatory bodies can achieve the goals of data protection law and significantly improve legal certainty by adopting quantified approaches and setting maximum re-identification risk thresholds to guide data privacy law and data protection law compliance in the context of biomedical data.[120]

---

117 Department for Digital, Culture, Media and Sport. Policy paper: Establishing a pro-innovation approach to regulating AI. Updated 20 July 2022. Available at: <https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement> accessed 27 February 2023; Department for Science, Innovation and Technology and Office for Artificial Intelligence. Policy Paper: AI regulation: a pro-innovation approach. 29th March 2023. Available at: <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach> accessed 27 February 2023

118 Purtova N. The law of everything. Broad concept of personal data and future of EU data protection law. Law, Innovation and Technology. 2018 Jan 2;10(1):40-81

119 Dalla Corte L. Scoping personal data: towards a nuanced interpretation of the material scope of EU data protecton law. European Journal of Law and Technology. 2019 May 16;10(1)

120 Bernier A, Knoppers B. Biomedical Data Identifiability in Canada and the European Union: From Risk Qualification to Risk Quantification?. SCRIPTed. 2021;18:4

Finally, the generation of synthetic data from patient data described in Section 2 of this report takes place in a well-governed context comprising a range of relevant laws and policies, as well as information governance standards and training requirements. For example, any use of confidential patient information in the health and care system is governed by the Caldicott Principles[121] and the use of patient data for research purposes will be subject to ethical review. Perhaps most importantly, data protection law will clearly apply in full to the processing of personal data involved in the development and training of synthetic data models.

The proper application of data protection principles, rights and obligations to this part of the process could go a very long way toward addressing concerns about harms or misuses of data arising from synthetic data generation. For example, the Art 9(2)(j) condition that should be met in order to process special category health data for research purposes requires additional safeguards under UK law, including that:

- processing must not be 'likely to cause substantial damage or substantial distress to a data subject',[122] and that;
- if processing is carried out for the purposes of 'measures or decisions with respect to a particular data subject' it must have been approved by a research ethics committee.[123]

Data protection law also requires that data subjects are informed about how their data is going to be processed and that such processing is fair. If the processing of patient data to generate synthetic data for purposes outside the reasonable expectations of the patients, that is likely to breach the first principle of data protection law.[124]

In short, it could be argued that the thorough application of data protection law, the common law of confidentiality, the Caldicott principles and relevant policies and codes governing patient data in the development of a synthetic data model safeguards individuals sufficiently without over-stretching the law to apply to fully synthetic output data.

---

121 National Data Guardian. The Caldicott Principles. 8 December 2020. Available at: <https://www.gov.uk/government/publications/the-caldicott-principles> accessed 27 February 2023

122 Data Protection Act 2018 (DPA), s19(2)

123 This includes research ethics committees and other bodies approved by the HRA, NHS organisations, the Secretary of State and some other sources of authority. DPA 2018 ss19(3)-(4)

124 Elizabeth Denham. Letter sent to: Sir David Sloman. 3rd July 2017. Available at: <https://ico.org.uk/media/action-weve-taken/undertakings/2014353/undertaking-cover-letter-revised-04072017-to-first-person.pdf> accessed 27 February 2023

Consideration: To ensure an appropriate balance is struck between privacy and health data innovation, policymakers, regulators and technical specialists should work together to consider a range of questions relating to the technical reality of synthetic data and the appropriate regulatory response to this form of data generation in light of the wider legal framework.

What might an alternative approach look like?

Developing an alternative approach to one which automatically views all forms of data synthesis as a form of anonymisation or privacy enhancing technology and sets a default that data will be considered personal data, will also require significant further debate and consideration. However, the position set out by the NHS Health Research Authority in its guidance on legal requirements for using health and care data provides one example. In a recent update to encompass synthetic data this now states:

> 'Synthetic data is neither personal data nor confidential patient information. It is not subject to data protection legislation or the common law duty of confidentiality.
>
> Where data is created artificially from confidential patient information or personal data, however, the act of creating it through a process of information synthesis is subject to the common law duty of confidentiality and data protection legislation - in the same way that the process of anonymisation is covered by these legal frameworks. See Section 2.2 above.
>
> Where synthetic data is generated to be statistically consistent with a real data set that it replaces, moreover, an assessment should be carried out regarding the likelihood of individuals being re-identified from the synthesised data. If necessary, additional safeguards may be needed to ensure that any reidentification risks (or other privacy risks) are sufficiently remote.'[125]

This guidance contrasts synthetic data with personal or confidential patient information. For this to be the case, a great deal turns on what is meant by synthetic data. However, the approach that the HRA has taken is to reverse the presumption adopted in the 'orthodox' approach to synthetic data. Beginning with a presumption that synthetic data are not personal data but cautioning that an assessment of the likelihood of individuals being re-identified may be required and that this could require further safeguards (although not going as far as suggesting that this could bring synthetic data back into the scope of 'personal data').

---

125 NHS Health Research Authority. Types of health and care information and the legal frameworks protecting them. Updated 5 Dec 2022. Available at: <https://www.hra.nhs.uk/planning-and-improving-research/research-planning/how-were-supporting-data-driven-technology/overview-legal-requirements-using-health-and-care-data-development-and-deployment-data-driven-technologies/2-types-health-and-care-information-and-legal-frameworks-protecting-them/> accessed 27 February 2023

While this guidance is deliberately brief, it is feasible that a shift in approach so that some forms of synthetic data are presumptively considered non-personal data would be appropriate. However, for data controllers, data users and patients/publics to have sufficient certainty and confidence that synthetic data are appropriately governed this requires further consideration.

In particular, technical, regulatory and policy experts should consider whether some forms of synthetic health data generation can be specified whereby the risk of identification is remote or negligible and therefore a reversal of the presumption could be justified? Our initial analysis suggests this could be more likely in the case of fully synthetic data generation using machine learning methods and technical measures in place to reduce overfitting, remove outliers and scrutinise for accidental matches in the output data.

Consideration: It should be considered whether a shift in presumption to viewing some forms of synthetic data as non-personal data, unless demonstrated otherwise, might be feasible for some clearly prescribed forms of synthetic health data generation.

Summary

Analysis of the legal framework, latest guidance and commentary identifies an approach and set of factors that will almost certainly be key to determining whether, or in what circumstances synthetic data will be considered personal data in the short-to-medium term. This views synthetic data generation primarily as a form of anonymisation or privacy-enhancing technology and begins by considering whether the source or training data are personal data. If this is the case, then a presumption is likely that data will remain personal data unless effective anonymisation can be demonstrated with confidence. To do so will require 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 data 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 appropriate as well as conducting penetration or motivated intruder testing. An audit or data protection impact assessment (see further detail in the Annex) 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. Nevertheless, 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.

However, there are challenges with this orthodox approach in the context of synthetic health data generation which may limit access to high-fidelity synthetic data for public interest purposes, such as the development of novel medical tools and for forms of health research that otherwise may not be feasible.

There may be reasons to adopt an alternative, more proportionate approach to the question of synthetic data as personal data. Some are technical questions and relate to the extent to which synthetic data and models genuinely incorporate or 'remember' personal data and the extent to which privacy risks therefore remain. Others are a matter of broader regulatory and policy decisions about the appropriate role of data protection, as opposed to other parts of the legal framework, governing synthetic data processing. If relevant potential harms are better regulated via other parts of the developing legal framework there may be less pressure on data protection law and encouragement to expand the scope of personal data.

Health data is a highly regulated arena and subject to a range of existing laws, policies and codes. Could this, combined with a thorough application of data protection law to the generation of synthetic data, provide sufficient reassurance that privacy, fairness and even public benefit are secured, without classifying all output data as personal data? A range of such questions need to be explored more fully by technical experts, regulators and policymakers to determine whether a shift in approach is appropriate.

However, it is possible that shifting the presumption that synthetic data are personal data for some forms of fully synthetic data generation (with controls for overfitting and removal of outliers for example) might be appropriate and defensible. However, securing confidence in this approach will require clear specification of which forms of synthetic data generation, in combination with which safeguards, could presumptively give rise to non-personal data.

Conclusions

- Recommendations

6. Conclusions

The term synthetic data encompasses a broad range of potential methodologies used to generate artificial data. This means that there is no one-size-fits-all answer to the question of whether synthetic data are 'personal data' under data protection law (the UK GDPR or the EU GDPR).

The innovative work of the CPRD and MHRA in this area provides an excellent example of the promise of synthetic data for health research and medical device development and testing purposes. It also highlights a difference in perspective to the data regulators and many commentators who currently view synthetic data primarily as a privacy enhancing technology or anonymization technique.

While some use cases for synthetic data relate to avoiding the burden of compliance that is in place for processing real patient data, there are other important reasons for developing synthetic health data including enabling developers to develop, test and validate the performance of medical devices.

Synthetic data is a rapidly developing field and we are at an early stage in the legal assessment of the privacy risks involved. Currently the authorities are sounding notes of cautious positivity about the potential of synthetic data to enhance privacy, but they recognise evidence of potential privacy risks. Neither authorities nor legal commentary to date are suggesting that synthetic data are not 'personal data'. An 'orthodox' approach is therefore most likely to be adopted by regulators assessing whether synthetic data are personal data.

This 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. Any assessment needs to be comprehensive and encompass both the data involved and the environment surrounding it, including organisational and legal safeguards in place. The challenge with this approach is that it may limit the development and availability of high-quality synthetic data for public interest purposes, such as health research and development of medical devices.

While a level of caution is to be expected in the case of any novel technology, there may 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) as central to data protection law. As the Government seeks to make changes to data protection law, adopting a proportionate approach which reduces the regulatory burden on some forms of AI-driven synthetic data generation may be consistent with broader regulatory principles.

There is an ongoing academic debate about whether the concept of 'personal data' has been overstretched, giving rise to ever greater complexity and uncertainty for data subjects and controllers about how to meet obligations and give effect to data subject rights. It could be that a more appropriate approach would 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 harm) that could arise from synthetic data generation.

Now would be good time for regulators (in particular the ICO), health data authorities, technical experts and legal specialists to come together and address whether the regulatory approach to synthetic health data generation is appropriate, or whether a different model, such as a shifted presumption that such data are non-personal data, may be more proportionate and technically feasible in certain clearly specified circumstances.

Recommendations

Throughout this report we have highlighted specific considerations for synthetic data developers, researchers, regulators or policymakers. On the basis of our analysis, 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.

Annex - Potential impact for Data Protection Impact Assessments (DPIA)

Annex - Potential impact for Data Protection Impact Assessments (DPIA)

In Section 5 we refer to the The UK Anonymisation Network's (UKAN) Anonymisation Decision Making Framework (ADF)[126] which provides a practical guide to assessing the identifiability of data and making decisions about safeguards and release models. There is also a more formal legal mechanism, the data protection impact assessment which is required under the UK GDPR (Article 35) in cases of processing that is likely to result in a high risk. This includes systematic and extensive profiling with significant effects, processing scale special category data (including health data) on a large scale and when using innovative technology with unknown and potentially novel risks.[127] The generation of synthetic health data could trigger several of these elements and it would therefore be best practice to conduct a DPIA to systematically and comprehensively analyse the processing and identify data protection risks.

On the 15th March 2023, ICO updated its Guidance on AI and Data Protection. The updates focus on clarifying the requirements for fairness in AI and deliver on a key ICO commitment to 'help organisations adopt new technologies while protecting people and vulnerable groups.'[128] Beyond the ICO updates, organisations such as the UK Anonymisation Network have provided guidance for practical and operational application. This Annex therefore draws out the key considerations from these sources which specifically focus on how to approach the assessment of data processing in AI systems. Existing detailed guidance for conducting a data protection impact assessment as part of data processing activities in general already exists as a first port of call.[129] This Annex provides a non-comprehensive outline of the sorts of considerations those undertaking DPIA assessments may wish to take into account.

---

126 UK Anonymisation Network (UKAN), Anonymisation Decision Making Framework (ADF). Available at: <https://ukanon.net/framework/> accessed 22 March 2023

127 Information Commissioner's Officer (ICO), 'Data Protection Impact Assessments (DPIAs)'. Available at: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias> accessed 22 March 2023

128 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

129 Information Commissioner's Office, Data Protection Impact Assessments (ICO, Updated October 2022). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/> accessed 22 March 2023

Consequently, the ICO advises that assessing data protection impact remains a context-based assessment whilst they recognise and emphasise that the law does not require a zero-tolerance approach to protecting data subjects' rights.[130] DPIA will still require the assessment of risks to individual rights, a plan for how these risks are to be addressed and an assessment of the impact this has on the use of this AI.[131]

When to undertake a DPIA for AI Processing and Synthetic Data?

Article 35(1) requires that an assessment of proposed processing activities is undertaken via a DPIA. DPIAs are essential exercises that result in a roadmap for data controllers, processors and the Regulator (ICO) to assess risks and to address and mitigate them to ensure data subject's rights are upheld. AI data processing often results in high risk to individual's rights and freedoms and consequently requires that a DPIA is undertaken.[132] If residual high risk is found the ICO requires further consultation before processing starts. In some contexts, other kinds of impact assessments may be required (or you may do so voluntarily) such as equality or algorithm impact assessments. Nevertheless, the ICO acknowledges that not all AI involves high risk processing but notes that, 'Article 35(3)(a) of the UK GDPR requires a DPIA where AI involves:

- systematic and extensive evaluation of personal aspects based on automated processing, including profiling, on which decisions are made that produce legal or similarly significant effects;
- large-scale processing of special categories of personal data;
- or systematic monitoring of publicly accessible areas on a large scale.'[133]

Consequently, the larger the scale of processing and if it is evaluative and systematic in nature, the more likely it will amount to high risk and require a DPIA. To help data controllers and processors, the ICO has produced a list of processing operations 'likely to result in high risk'.[134] Therefore, DPIAs amount to a holistic assessment where assessing if processing is high risk will require looking at how the risk in its individual elements impact the overall risk of the system.

---

130 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023; Data Protection Act 2018, Section 64(4)

131 Data Protection Act 2018, Section 64(3)

132 Data Protection Act 2018, Section 64(1)

133 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

134 Information Commissioner's Office, Guide to Data Protection: Examples of Processing Likely to Result in High Risk' (March 2018) Available at: <https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/data-protection-impact-assessments-dpias-1-0.pdf> accessed 22 March 2023

The ICO Guidance on AI does mention synthetic data. It states that, 'to the extent that synthetic data cannot be related to identified or identifiable living individuals, it is not personal data and therefore data protection obligations do not apply when you process it.'[135] However, the Guidance states that, 'you will generally need some real data which was used to estimate those realistic parameters, by analysing the synthetic data', the risk of re-identification or the amount of obfuscation needed in the de-identification processes may be too challenging to rule out that such data is personal data.

As such, it seems for now at least that the ICO views that most methods of synthetic data generation will either rely on real data to create the model, use it to continue to evaluate its utility, or could possibly reidentify an individual bringing in it within the scope of the GDPR. As such, a cautionary approach seems sensible in deciding whether to undertake a DPIA.

Key point:

Whether a DPIA is necessary is still context-dependent whether or not AI processing is involved. However, the UK GDPR Article 35(3)(a) and Data Protection Act 2018, Section 64(1) require a DPIA for certain types of AI processing that they categorise as high risk.

What should your DPIA assess?

DPIAs outline the nature, scope, context and purposes of processing personal data. The purpose is to make clear how and why you are using AI to process the data. The ICO advises that you will need to detail factors such as:

- how to collect, store and use data;
- the volume, variety and sensitivity of the data;
- the nature of your relationship with individuals; and
- the intended outcomes for individuals or wider society, as well as for you;
- what additional measures you plan to take;
- whether each risk has been eliminated, reduced or accepted;
- the overall level of 'residual risk' after taking additional measures;
- the opinion of your DPO, if you have one; and whether you need to consult the ICO.[136]

---

135 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

136 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

Non-exhaustive considerations for AI

Evaluation is context-dependent and evidence that less risky alternatives that achieve the same purpose of the processing must be considered and reasons given for why they were not alternatively adopted. Additionally, impact considerations should take account of both allocative and representational harm, that is harm that impacts resources and also equal treatment of different groups.

They also suggest that when describing processing that a systematic description should be provided and an explanation of relevant variations or margins of error. This includes a description of the scope and context of processing, the data to be processed, the number of data subjects involved, the source of the data and the extent to which individuals are likely to expect processing. For automated decisions DPIAs should identify and record the degree of human involvement and where human intervention is envisioned, evidence of processes that ensure this is meaningful are important.

It is acknowledged that descriptions of AI processes are difficult but remain a necessary part of DPIAs. The ICO therefore advises that two versions should be kept; one for specialist audiences and another amounting to a high-level description explaining how personal data inputs affect data subjects.

DPIAs should also set out your role and obligations as a controller and include any processors involved. If AI systems are wholly or partly outsourced to external providers, both you and any other organisations involved should assess whether joint controllership exists under Article 26 of the UK GDPR and consequently collaborate on the DPIA process as needed. Controllership is particularly important for AI systems as it is commonplace that several organisations will be involved in developing and deploying AI systems which process personal data. Further guidance has been provided by the ICO on how to identify whether you amount to a controller or processor.[137]

Necessity and proportionality are also important principles for AI systems as their deployment must be 'driven by evidence that there is a problem, and a reasoned argument that AI is a sensible solution to that problem.'[138] They will require an assessment of the interests of using AI against its risks to the rights and freedoms of individuals that are guaranteed under data protection law. It is important to be aware that potentially impacted rights can go beyond individual data subjects being impacted to whole groups, as well as going beyond data protection law into other areas of legislation such as equality and discrimination.

---

137 Information Commissioner's Office, Guide to Data Protection: Controller and Processors (Updated October 2022) Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/controllers-and-processors/> accessed 22 March 2023

138 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

It is important to seek your data protection officer's advice at the outset of designing such systems as DPIAs will require that measures are identified to reduce and mitigate identified harm/ impacts. It may be the case that for some risks, no mitigating measures have been identified. Such residual risks should be documented on the DPIA and where residual risk remains high, the ICO requires that they are consulted before processing commences. Moreover, DPIAs are living documents meaning that although they are to be carried out before processing begins, they should be regularly reviewed and reassessed where appropriate. "Appropriate circumstances" would include if the scope, context, purpose, risk to individuals or nature of processing changes for any reason.

Training AI for the generation of synthetic data inevitably involves a trade-off between reducing the quantity of personal data used to train that system and training a sufficiently accurate AI system. In such circumstances it will be important to consider the balance between data minimisation and statistical accuracy.[139] Such trade-offs entail consideration of key data protection principles such as fairness, proportionality and accuracy. It is important to note that the ICO Guidance distinguishes between the principle of accuracy and statistical accuracy: statistical accuracy refers to the answers the AI gets correct or incorrect, the principle of accuracy amounts to a duty to ensure that personal data is not, 'incorrect or misleading as to any matter or fact, and where necessary, is corrected or deleted without delay.'[140] One therefore focuses on the statistical utility, and the other is a focus on accurate representation of the data subject and how inaccurate representation can detrimentally impact them.

Storage limitation is also important over the life cycle of an AI system to ensure fairness. Consequently, data must not be held longer than is necessary to achieve your purpose. This should not mean that AI systems cannot process personal data but instead should be viewed as a requirement to be transparent about who is adequate, necessary and proportionate to achieve the desired purpose.

A further point of consideration is security to protect the data held from unauthorised or unlawful processing, loss, destruction or damage. Recital 71 sets out the technical and organisational measures needed. The UK GDPR's security requirements apply to both the data you process and the systems and services you use for processing. There is no 'one-size-fits-all' approach to security. However, AI introduces new risks such as adversarial attacks because of factors such as dependence on third party code relationships and are often integrated with both new and existing IT components.

AI systems therefore function within a larger chain of software components, data flows, organisational workflows and business processes. As such, a systems-wide approach to security needs to be envisioned. Organisation processes and procedures for security need to account for the fact that a wider-array of professionals are likely to access

139 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

140 Information Commissioner's Office, Guidance on AI and Data Protection (updated 15 March 2023). Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/> accessed 22 March 2023

these systems within a single organisation and that they may have different levels of security hygiene and knowledge.

The ICO advises that it is not possible to list all known security risks but that the impact of AI on security depends on factors such as:

- the way the technology is built and deployed;
- the complexity of the organisation deploying it;
- the strength and maturity of the existing risk management capabilities; and
- the nature, scope, context and purposes of the processing of personal data by the AI system, and the risks posed to individuals as a result.[141]

Demonstrating appropriate accountability

To further assist, the ICO has created an Accountability Framework which is divided into 10 categories to help data controllers and processors understand how accountability is assessed. The aim here is not to provide a comprehensive list, as stated previously, the general theme is that each must be assessed on a case-by-case or context-dependent basis. The general themes under consideration are:

Accountability categories	Factors may include but are context-dependent
Leadership and oversight	Organisational structure, whether to appoint a DPO, appropriate reporting, operational roles, group to provide oversight and direction, and operational group meetings
Policies and procedures	Direction and support for staff on their roles and responsibilities, review and approval process to check policies and procedures and consistent and fit for purpose, staff awareness of relevant information governance policies and procedures for their role; and the adoption of a data by design and by default approach across the organisation's policies and procedures
Training and awareness	All staff training programme, induction and refresher training, specialised roles, monitoring, and awareness-raising

141 Information Commissioner's Office (ICO), Guide to Data Protection: Guidance on AI and Data Protection- How Should We Assess Security and Data Minimisation in AI? (updated March 2023) Available at: <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/#whatsecurityrisks> accessed 22 March 2023

Accountability categories	Factors may include but are context-dependent
Individual's rights	Informing individuals and identifying requests, resources, logging and tracking requests, timely responses, monitoring and evaluating performance, inaccurate or incomplete information, erasure, restriction, data portability, rights relating to automated decision-making and profiling, and individual complaints
Transparency	Privacy notice content, timely privacy information, effective privacy information, automated decision-making and profiling, staff awareness, privacy information review, tools supporting transparency and control
Records of processing and lawful basis	Data-mapping, records of processing activities (ROPA), ROPA requirements, good practice for ROPAs, documenting your lawful basis, lawful basis transparency, consent requirements, reviewing consent, risk-based age checks and parental/ guardian consent, and legitimate interest assessment (LIA)
Contracts and data sharing	Data sharing policies and procedures, data sharing agreements, restricted transfers, processors, controller-processor contract requirements, processor due diligence checks, processor compliance reviews, third-party products and services, and purpose limitations
Risks and data protection impact assessments (DPIAs)	Proper planning, mitigation and addressing of risk, evidencing that the least risky methods of processing used etc.

Accountability categories	Factors may include but are context-dependent
Records management and security	Creating, locating and retrieving records; secure data transfers, data quality, retention schedule, information asset register, destruction, rules for acceptable software use, access control, unauthorised access, remote working, secure areas; and business continuity, disaster recovery and backups
Breach response and monitoring	Detecting, managing and recording incidents and breaches; assessing and reporting breaches, notifying individuals, reviewing and monitoring, external audit or compliance check, internal audit programme, performance and compliance information; and use of management information

These categories provide an overview of the sorts of considerations that will need to be accounted for in the overall design of a system that incorporates AI.

Key points

- evaluation will still remain context-dependent. Depending on how the synthetic data was generated, how the AI is set up and managed, what other IT systems it operates in etc., will all add further considerations for identifiability and for adverse event management.
- the ICO has issued some AI specific guidance which may help but largely reiterates current guidance just with more commentary from an AI-specific environment.
- trade-offs for data minimisation and statistical accuracy will have relevance for some synthetic datasets. It is important to bear in mind the key principles of fairness, proportionality and accuracy when determining the line in such trade-offs.
- note the ICO differentiates the principle of accuracy with the need for statistical accuracy: one focuses on the impact on a data subject and the second focuses on the utility of the data.

UK Anonymisation Network guidance[142]

On the basis that in some situations synthetic data is still considered to be an anonymisation process or PET, the UK Anonymisation Network's guidance is highly relevant. These resources are more focused on operational use than ICO guidance. As this report is focussed on whether synthetic data could amount to personal data, their guidance that specifically focuses on the legal context (GDPR) may be of particular use to those undertaking DPIAs.

The purpose of these resources is to get data processors and controllers to think more broadly about what they are doing, why they are doing it, to know the ins and outs of their data sharing processes and capture when in the processing chain there could be risks to data subjects' rights and to mitigate them.

They therefore suggest 10 considerations across the life cycle of your data.[143]

1. To know your data/intended use (describe/ capture the presenting problem)

This task requires you to understand the 'data situation' amounting to a top-level description of what you intend to do. Relevant legal considerations will be Articles 6, 9, 5(1)(b), 5(2), 24 and 28 (and possibly others). These articles will require you to consider the lawfulness and fairness of processing and consequently what responsibilities you might hold. The purpose of this exercise is to understand the data before you try to anonymise it. Questions such as are there outliers, sensitive information, qualitative or quantitative, special cases and what combinations of variables exist. It will be important to identify which variables pose a risk to creating safe data.

2. Sketch the data flow

The process of sketching out the data flow from its point of origin to end will further help define your responsibilities. In some scenarios that data may be relatively straightforward, in others it may involve international data sharing. Moving data across multiple environments can be tricky for deciphering responsibilities, this is particularly the case with AI where parts of the system may be "managed" by third parties etc. It is not enough to rely on the class of data, i.e., personal data means X responsibility exists. Even where your role is downstream i.e., not data collection, that does not necessarily mean the origin of the data has little to do with you or that if you do not have access to personal data, you do not hold data controller responsibilities. If you are not the data controller, you are likely to need instructions from the controller on how you are to process it downstream, this still requires an overview understanding of the data's lifecycle. Likewise, even if you do not have access to personal data, if you are determining the means and/or purpose of processing you still may amount to a data controller.

---

142 UK Anonymisation Network (UKAN), The Anonymisation Decision Making Framework. Available at: <https://ukanon.net/framework/> accessed 27 March 2023

143 Mark Elliot, Elaine Mackey and Kieron O'Hara, 'The Anonymisation Decision Making Framework 2nd Edition: European Legal Context (GDPR)' (UKAN, October 2020). Available at: <https://ukanon.net/framework/> accessed 27 March 2023

3. Map the data flow environment

Once you are aware of the data flow, you will need to assess the agents, other data, governance and infrastructure along its path. This will allow you to assess risks for identification, attacks or unlawful use or access. The environment is not specifically addressed by the GDPR but it does provide guidance on the appropriate organisational and technical measures to ensure a secure environment. Relevant articles include 5(1)(f) and Recital 26 (means reasonably likely) among possible others. For example, synthetic data, derived from patient data, could have a high risk of reidentification when used by other users who have access to large patient data stores. Such considerations will be relevant when mapping the environment.

4. Describe the data and map risk

Within each environment you will need to describe the data such as its structure, type, variable type, population, topic sensitivity etc and then use these parameters to map risk, both in terms of the likelihood or impact of a given breach.

Certain types of data are categorised as “special category” due to the risk they inherently pose as identifying or where a breach could have a particularly harmful impact on a data subject’s rights and interests protected in law. Additionally, some additional types of data have been discussed as potentially amounting to personal data in subsequent WP or CJEU guidance such as online identifiers, genetic and location data. Relevant articles might include Articles 9, 6 and 10.

5. Ethical requirement for engagement with stakeholders

Realistically, the data situation will often mean that zero risk is unrealistic, even with anonymised data and it is therefore good practice to engage with relevant stakeholders to ensure that the purpose and way in which data is to be processed is expected. This will also help meet the legal requirements to ensure that processing is fair and transparent. Relevant articles of the GDPR might include Article 6, Recital 50, Articles 12-14, 34, 35(9) and 36.

6. Evaluate risk

Once the data flow has been mapped out, stakeholders and roles identified, including properties of the data and environment you should be able to map out risk. Chapter 4 of the GDPR, in particular, Articles 25, 32, 35 and 35(7) provide further details on what needs to be covered. The ultimate purpose of this exercise is to assess if residual risk exists and if so, how it will be mitigated or removed. Risk assessment for DPIAs is an iterative process and will require reassessment, particularly if you change purposes for processing the data among other triggering reasons.

7. Implementation of testing for risk and controls for disclosure risk

At this stage, some residual risk may be high. You will therefore need to select methods, proportionate to the risk to assess it for example intruder testing or data analytical risk assessment. Alternatively, you may decide to add controls to mitigate that risk such as differential privacy or k-anonymity models. These exercises are not just about reducing the likelihood of these adverse events but also reducing the impact/ harm should they occur.

8. Stakeholders' trust

Trust is not just about behaving in a trustworthy manner but engaging in a meaningful way. That means that communication must be transparent to ensure that stakeholders are aware of relevant changes and that they have a point of contact to communicate concerns. This is a key consideration because how much freedom you have to address an adverse event will likely be dependent on how stakeholders trust you.

9. Crisis management

As this is a risk management exercise you should still consider what will happen if an adverse event occurs. For example, what crisis management policies are in place and how will these events be addressed. There are some legal requirements that must be met in the event of a breach such as Article 33(1) which stipulates when breaches need to be reported and if you need to notify supervisory bodies. Additionally, Article 34 requires communication with the data subjects where a breach results in high risk to the rights and freedoms of data subjects. There are exemptions for Article 34 such as if encryption means that breached data is unintelligible or subsequent measures to ensure high risk is an unlikely outcome or notification would involve disproportionate effort.

10. Ongoing surveillance

Data and particularly data used in AI systems is often dynamic, meaning that risk is constantly evolving. As such, procedures will need to be in place to monitor your data situation and environment. Moreover, anonymised data may be functionally anonymised for processors or end users but still be considered personal data for the data controller. Moreover, responsibility is still held by the controller to ensure data remain anonymised and that risks for identification or other adverse events are assessed and managed appropriately. Hence DPIAs are considered living documents.

Key points

- synthetic datasets that start by processing personal data will still involve ongoing risk assessment and mitigation duties for data controllers and possibly some for processors and data users depending on the data situation and environment
- as with any innovation, there may be risks that should be assessed in much the same way as if they were considered personal data not because a legal requirement exists but as good practice and to merit trustworthiness.
- as with any risk assessment, there is not a prescriptive approach, and a case-by-case assessment will need to be undertaken to understand synthetic data's data situation and the wider environment it operates in.

The PHG Foundation is a non-profit think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare and deliver improvements in health for patients and citizens.

intelligence@phgfoundation.org

UNIVERSITY OF CAMBRIDGE

PHG FOUNDATION

making science work for health