



UK Synthetic Data
Community Group

Synthetic Data Release Framework under UK Data Protection Law

A Decision-Making Framework for
Understanding Synthetic Data Release
from Trusted Research Environments
under UK Data Protection Laws



DARE UK



DARE UK
COMMUNITY GROUPS



UK Synthetic Data Community Group

This non-statutory good practice guidance has been developed by the UK Synthetic Data Community Group, informed by expert discussion and practical experience. It is intended to help organisations assess the generation, evaluation and release of synthetic datasets under UK data protection law. It does not cover governance relating to the release of generative models, which may also carry separate disclosure risks. Nor does it constitute ICO guidance, legal advice, or a binding regulatory code. Organisations remain responsible for their own legal and governance decisions. References in this document to UK GDPR are limited to the UK GDPR as amended by the Data (Use and Access) Act 2025 as in force at the date of publication. Where relevant, organisations should also consider the Data Protection Act 2018, confidentiality obligations, contractual controls, intellectual property, and sector-specific law.

This document is an implementation-oriented companion to the UK Synthetic Data Community Group's VSTAR framework, providing more detailed governance and release guidance for synthetic datasets under UK data protection law. The following sections summarise the key considerations during the various phases leading up to and after synthetic data release. Five phases are considered: design, generation, evaluation, governance and review, post-release monitoring.

Funded by



DARE UK



DARE UK

COMMUNITY GROUPS



**UK Synthetic Data
Community Group**

Authors & Contributors



Anmol Arora
Academic Clinical
Fellow



Lewis Hotchkiss
Senior Research Officer



Puja Myles
Director of CPRD



Emily Oliver
Head of Research &
Capacity Building



Cristina Magder
Head of Collection
Development



Paul Comerford
Principal Technology
Policy Advisor



Maddy Griffiths
Senior Policy Officer



Sophie McCall
Senior Data Analyst



Gillian Raab
Research Fellow



May Yong
Senior Research
Software Engineer



Zoya Yasmine
DPhil Student



Mark Elliot
Professor



Peter Wright
Head of Information
Governance



Colin Mitchell
Head of Humanities



Elizabeth Redrup-Hill
Regulatory Policy Lead



Sharon Heys
Head of Legal



Markus Trengove
Senior AI Policy
Manager



Cassie Smith
Director of Legal, Trust
& Ethics



Zuzanna Domaradzka



Ryan Bremner
Information
Governance Officer

What is the Purpose of this Framework?

The primary objective of this framework is to provide non-statutory, good practice guidance for organisations navigating the complexities of generating, evaluating, and releasing synthetic datasets under UK data protection law. This initiative was born out of direct engagement by the UK Synthetic Data Community Group with various data owners, which revealed a significant knowledge gap. Stakeholders identified that specific guidance regarding synthetic data under UK GDPR was a crucial requirement to help them understand how to make consistent, defensible decisions regarding the release of these datasets.

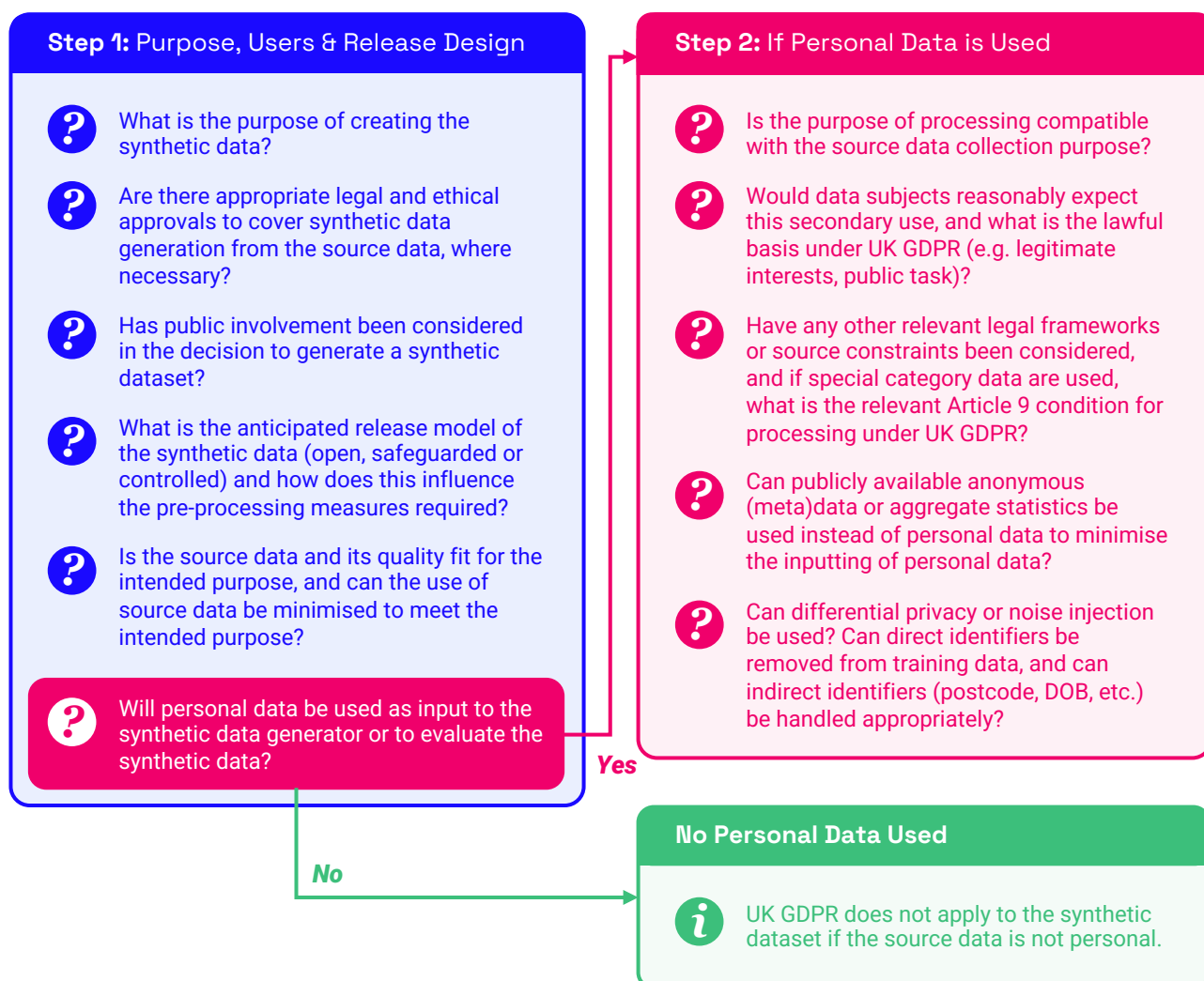
In response to this need, the Community Group hosted two intensive workshops designed to bridge the gap between technical possibility and regulatory compliance. These sessions brought together a diverse cohort of synthetic data experts, legal professionals, regulatory authorities, and information governance specialists to develop and refine a practical decision-making structure. By synthesising these various perspectives, the framework offers a holistic view of the risks and responsibilities involved, including legal and ethical considerations.

The framework is structured around the complete synthetic data lifecycle, providing a step-by-step roadmap for organisations to move from initial concept to post-release oversight. It begins with Phase 1: The Design Stage, where data owners define the intended use and release model, and Phase 2: The Generation Stage, which focuses on method selection and the necessity of a Data Protection Impact Assessment (DPIA). This is followed by Phase 3: The Evaluation Stage, involving rigorous assessment of fidelity, utility, and disclosure risk using tests like the "motivated intruder test," and Phase 4: Governance & Review, where a formal legal determination is made regarding the data's status under UK GDPR. The lifecycle concludes with Phase 5: Post-Release Monitoring, ensuring that datasets are regularly reviewed against emerging re-identification threats or evidence of misuse.

Ultimately, it is hoped that this framework serves as an essential, implementation-oriented companion to the VSTAR framework. By outlining necessary steps throughout the entire synthetic data lifecycle - including design, generation, evaluation, governance, and post-release monitoring - it enables data owners to ensure they understand and meet the requirements of the UK GDPR (as amended by the Data (Use and Access) Act 2025). Through this structured approach, the framework aims to build trust in synthetic data as a privacy-enhancing technology while ensuring organisations remain responsible for their own legal and governance decisions.

Phase 1: The Design Stage

At the outset, organisations should define the intended use, expected users, and release model for the synthetic data, because these decisions shape both the appropriate fidelity level and the assessment of identifiability. Where personal data are used as inputs, organisations should separately assess the lawfulness of that upstream processing and the legal status of the synthetic output in its intended release context. Using metadata or aggregate statistics instead of record-level personal data may reduce disclosure risk but does not by itself determine whether the output is anonymous. Likewise, the fact that source data are publicly available does not on its own mean that the synthetic output can be released publicly; that depends on identifiability, other legal or contractual restrictions, and the chosen release model. If the synthetic data will be anonymous at the point of release, creating it from personal data will usually fit within the original purpose for collecting that data. Even so, there can be exceptions, so each case should still be assessed individually (ICO Anonymisation Guidance).



Phase 2: The Generation Stage

Where personal data are used to generate synthetic data, organisations should assess at an early stage whether the processing is likely to result in a high risk to individuals' rights and freedoms and therefore requires a DPIA. In many synthetic data projects, a DPIA will be advisable or required because the processing may involve innovative technologies, special category data, large-scale processing, or data matching. Some organisations may wish to complete a DPIA even when this is not required by UK GDPR. The decision on which model to use when generating synthetic data is dependent upon the format of the source data available (e.g. raw data or aggregate statistics) to train the generator and the organisation's choice of fidelity level of the synthetic data to be generated. Generating high-fidelity data may benefit from a machine learning approach, though these are typically less explainable. Generating low-fidelity data may simply require rule-based approaches that are easily explainable, but which are unlikely to replicate unknown or especially complex correlational patterns. Open-source generators are easier to investigate if any output datasets are subsequently challenged. Proprietary algorithms are still permissible, but organisations should consider seeking independent verification of performance and consider wider factors that are likely to affect decision-making such as cost and vendor lock-in.

Step 3: Documentation & DPIA

- ❓ Does the processing trigger the ICO's DPIA high-risk criteria, particularly because it involves sensitive or highly personal data, or innovative technological or organisational solutions? (ICO, *Examples of processing 'likely to result in high risk', 2023*)
 - ➔ ❓ If yes, complete a DPIA before the relevant personal-data processing begins.
- ❓ What harms could arise from leakage, memorisation, singling out, linkage, inference, misuse or misleading downstream use?
- ❓ What mitigations reduce those risks, and who approved them?

Step 4: Method Selection

- ❓ What specific generation approach is being used, and how explainable is it in practice?
- ❓ Is the method of generating synthetic data open-source or proprietary?
 - ➔ ❓ If the method is proprietary, does the vendor provide sufficient information about the algorithms used and its testing (e.g. identifiability, bias, etc) to enable independent verification?
- ❓ What safeguards could be used to reduce verbatim or near-verbatim reproduction of source records?
- ❓ How have protected characteristics been handled during the synthetic data generation and are there likely biases that users should be aware of?

Phase 3: The Evaluation Stage

Broadly, organisations may wish to assess the fidelity, utility and privacy of synthetic data. Fidelity refers to how closely the synthetic data resembles the original dataset. High-fidelity data can support analyses, but it also increases disclosure risk as more detail from the source is preserved. However, lower fidelity does not automatically confer lower disclosure risk because differences from the original at an aggregate or population level can still contain patterns that allow individuals to be re-identified. Utility is a measure of how useful the data is for its intended use and is a subjective analysis for each use case. Disclosure risk concerns whether individuals can be singled out, linked, or have attributes inferred from the synthetic data, alone or in combination with other information reasonably likely to be available in the release context considering cost, time, available technology and likely technological developments. If the disclosing controller cannot identify anyone from the data to be released, even when taking account of the source data it holds and means reasonably likely to be used, that is a strong indicator of anonymity; however, disclosure risk must still be assessed in the hands of the intended recipient as being sufficiently remote. Where the data are being disclosed as anonymous information to an independent recipient, disclosure risk should be considered using the UK 'motivated intruder test'.

Disclosure risk metrics are particularly important when organisations have generated synthetic data using non-explainable algorithms that have been directly exposed to the original source data. In cases where the synthetic data generator has only been exposed to publicly sourced anonymous (meta)data or aggregate statistics, post-hoc evaluation metrics are less important for the evaluation of disclosure risk. There is no consensus within academic or industry on what disclosure risk metrics should be used to evaluate synthetic data. We recommend that organisations use more than one metric. There are metrics that attempt to proxy aspects of identifiability risk, such as singling out, linkability and inference, but they cannot be relied upon in isolation to determine whether data are anonymous under UK GDPR. Disclosure risk metrics and statistical tests should be treated as evidential tools rather than definitive proof of anonymity. No single metric can, in isolation, demonstrate that synthetic data is not personal under UK GDPR and metrics should be interpreted in light of the data generation methodology and context of release.

Step 5: Assess Identifiability in the Intended Release Context

- ❓ In the intended release setting, could an individual be identified, singled out, linked to other data or inferred with means reasonably likely to be used?
- ❓ Has identifiability been assessed by reference to the release model, recipient group, access conditions, and other data reasonably likely to be available?
- ❓ Where the data are being disclosed as anonymous information to an independent recipient, consider whether a 'motivated intruder test' would be helpful in assessing the disclosure risk? (*ICO, How do we ensure anonymisation is effective? 2025*)
- ❓ What external datasets, background knowledge, or public information are reasonably likely to be available to recipients or adversaries?
- ❓ How have unique, rare, or outlier records been treated?
- ❓ When assessing disclosure risk, consider who will receive and use the synthetic data. The ICO's "whose hands?" test applies to independent recipients, but not to processors or joint controllers (*ICO, How do we ensure anonymisation is effective? 2025*).
- ❓ Consider how much information about the synthetic data generation model is required to make an informed assessment of the privacy risks.

Step 6: Ethical & Bias Review






- ❓ What are the intended benefits of releasing the synthetic data and do these outweigh the potential harms?
- ❓ Have any assumptions been made about individuals within the source data when generating the synthetic data? If so, what?
- ❓ Could downstream users infer sensitive characteristics, apply the data outside its intended purpose, or overstate its validity?
 - ↳ ❓ If yes, has the intended use and limitations been clearly stated (e.g. in an end-user license agreement)?
- ❓ Could outputs cause discrimination or unfair outcomes?
 - ↳ ❓ If yes, has the intended use and limitations been clearly stated?

Phase 4: Governance & Review





Following evaluation, organisations should make and document a formal determination as to whether the synthetic data constitutes personal data under UK GDPR, taking account of identifiability, means reasonably likely to be used, and the context of release for intended recipients. When assessing anonymity, organisations should consider the extent to which individuals could be singled out, linked across datasets, or subject to inference, taking account of means reasonably likely to be used. If the synthetic data is not personal under UK GDPR, organisations may choose how to release the data depending on organisational, contractual and intellectual property factors. If the data is subject to UK GDPR, they must in addition treat the data as personal data and consider relevant articles, including in relation to legal bases and Data Protection Impact Assessments (DPIAs). The release of data that is anonymous in the relevant hands does not itself require a DPIA under UK GDPR. However, where personal data were used upstream to generate the synthetic data, organisations should separately assess whether that earlier processing requires a DPIA. Organisations should keep written records of their decision-making process for evaluating and releasing the synthetic data. Organisations may wish to consider preparing End User License Agreements regardless of their access model. Organisations should publish enough information to enable informed scrutiny of the synthetic dataset's provenance, intended use, fidelity, release model, and evaluation approach, while withholding technical detail that would materially increase disclosure or attack risk.

Decision: Does the released synthetic dataset constitute personal data under UK GDPR in the chosen release setting?




Not Personal Data

-  UK GDPR does not apply to the released synthetic dataset.
-  Choose a release model subject to governance, confidentiality, contractual, IP and sector-specific limits.
-  Open release may be possible if residual risk is acceptable.
-  Keep records, label the data clearly as synthetic, and define acceptable uses.
-  Risk assessment should be completed, though not necessarily a DPIA.

Personal Data

-  Treat the released synthetic dataset as personal data under UK GDPR.
-  Identify relevant UK GDPR articles, lawful basis for each processing activity, and whether a DPIA is required or advisable.
-  Use appropriate access controls, agreements and governance safeguards.
-  Keep records, publish enough for scrutiny, but withhold details that would materially increase attack risk.

Step 7: Third-Party Risk & Data Sharing Controls

-  Could specific recipients later gain access to external datasets which could increase the disclosure risk of the synthetic data?
-  Is an access agreement needed to clarify what would constitute 'misuse' of the synthetic data?
-  Are reasonable technical and organisational measures in place to mitigate the risk of misuse without limiting intended utility (e.g. contractual usage restrictions)?

Step 8: Transparency & Accountability Review

- ❓ Publicly disclose enough information about the intended use of the synthetic data, the nature of the training data, the type of generation method used, evaluation techniques used and the determination of the dataset's legal status to enable informed scrutiny without materially increasing disclosure risk.
- ❓ Is the synthetic data clearly labelled as synthetic, with its intended fidelity level and appropriate use limitations?

Step 9: Record-Keeping

- ❓ Privately record details of the model and detailed risk assessments that could materially increase disclosure risk.
- ❓ Is there an audit trail of the governance process for releasing the synthetic dataset?

Phase 5: Post-Release Monitoring

Organisations should be aware that the extent to which data can be re-identified is ever-changing and they should keep synthetic datasets under regular review in case changes to the access model are required. It is possible that data previously considered to be anonymous data may, due to technical advances in future, become potentially identifiable and organisations should be proactive in their response to emerging developments. Organisations should also review how their synthetic data is being used and by whom, in order to consider whether it falls within the intended use e.g. if a low-fidelity synthetic dataset is being misused to generate research findings when it was only intended for teaching and training, the organisation may wish to review whether its dissemination practices remain appropriate. Organisations should consider specific review triggers for re-assessing synthetic datasets, such as the emergence of new external datasets, advances in re-identification techniques or evidence of misuse. Where risks increase, organisations should consider modifying access controls, withdrawing the dataset, or reclassifying its legal status. In the event of a breach, organisations should promptly contain and assess the incident and comply with any applicable notification obligations.

Step 10: Ongoing Risk Monitoring



Is there a process in place to check if new attack vectors, vulnerabilities or newly released external datasets increase disclosure risk over time?



Monitor misuse and purpose creep beyond the acceptable uses of the synthetic data.



Is there a process for dataset withdrawal if risks emerge?



Arrange for appropriate version-control to limit deprecation.

Frequently Asked Questions

This FAQ is intended to support risk-based decision-making and does not provide absolute guarantees or legal advice. Organisations remain responsible for assessing synthetic data in context and documenting their decisions.

In what cases is synthetic data useful?

This framework focuses on the use of synthetic data as a privacy-enhancing technology. Synthetic data may be released publicly or restricted in its release. Publicly released synthetic data can be useful to help students, researchers and the public understand the structure of data that organisations hold. It may also be useful for researchers to test code and conduct feasibility tests on what analyses are possible using data that organisations hold. These use cases generally require low-fidelity synthetic data which mirrors the structure of the source data without holding any patterns that could be used for generating conclusions. Organisations may also wish to create higher-fidelity synthetic data which could be used for analysis, generating conclusions or producing real-world predictions but these synthetic datasets would generally carry a higher risk of people being re-identified from the synthetic data, so would often have some safeguards in their release. An emerging use case of synthetic data is for data holders who typically make confidential data available through secure data environments (SDEs) or trusted research environments (TREs) but wish to use synthetic data to more openly showcase the data they hold within the TRE.

Should synthetic data only be released within a Trusted Research Environment (TRE)?

There is no legal requirement to use a TRE for any form of synthetic data. However, this may be useful if it is considered personal data in the hands of recipients or because recipients may have means reasonably likely to link it back to real individuals. In many circumstances, synthetic data may be suitable for wider release, subject to a careful assessment of residual risk and appropriate governance controls.

Is there a recommended disclosure risk metric and quantitative threshold for what constitutes anonymous synthetic data?

There are no prescribed thresholds for what constitutes personal data or not. This is because depending on the synthetic data to be evaluated, it may be preferable to evaluate the methodology used to create it; the scores achieved with different metrics or a combination of the two approaches. Metrics are a useful indicator of risk, but organisations should focus on context-dependent tests for anonymity under UK GDPR, such as the 'motivated intruder test' and 'means reasonably likely to be used' test. There is no guarantee that currently available metrics will be able to account for disclosure risks of the future and so in general, we advise focussing on the source data and methodology used to create the synthetic data rather than sole reliance upon post-hoc evaluation by metrics. Under UK GDPR, organisations do not need to take into account purely hypothetical or theoretical chances of identifiability but should assess whether identification is reasonably likely in light of objective factors such as cost, time, available technology, and foreseeable developments.

Once the synthetic data is generated, if records identical to those in the source dataset are generated by chance should they be removed?

Any coincidental generation of a record which is identical to or even very similar to a record in the source dataset should trigger review but does not automatically indicate concern. The appropriate response depends on the release model, the uniqueness of the record, whether it appears to correspond to a person in the source data, and whether checks against the source data or source statistics indicate a material increase in singling-out, linkage, or inference risk. Possible mitigations include alteration, suppression, regeneration, or tighter access controls. There is no single universal rule, and organisations should document their reasoning.

What are the legal considerations of synthetic data?

Synthetic data may or may not be personal data, depending on how it is defined, how it was generated, its anticipated use and whether individuals present in the source data can be identified from the data. If synthetic data is not personal data, then UK GDPR does not apply, and it can be shared without data protection restrictions subject to appropriate assessment and governance. However, organisations may still be cautious due to reputational or intellectual property concerns. If the synthetic data is assessed as falling under the definition of personal data, then it is still within scope of UK GDPR and must be treated as if it were personal data. It should be noted that even when synthetic data is not personal data, this does not permit unrestricted lawful use of the data as there are many different types of sector-specific legislation that limit the use of data analytics regardless of the type of data, e.g. for personalisation or targeting. Election law, consumer protection, intellectual property, financial services, health research, advertising law may all apply regardless of identifiability of training data.

Do I need a legal basis under UK GDPR to create synthetic data?

If personal data is being processed to generate the synthetic data a legal basis will always be required, even if the synthetic data being generated is anonymous. If the synthetic data will be anonymous at the point of release, creating it from personal data will often fit within the original purpose for collecting that data. Even so, there can be exceptions, so each case should still be assessed individually (ICO Anonymisation Guidance).

Do I need a legal basis under UK GDPR to process synthetic data which is considered anonymous data?

If the synthetic data is genuinely anonymous in the hands of the person or organisation processing it, UK GDPR does not require a legal basis for that processing. However, the same dataset may still be personal data in another party's hands if they hold additional information or have means reasonably likely to identify individuals, or if the data relates to an identified or identifiable person.

Do I need to do a Data Protection Impact Assessment before releasing synthetic data?

Article 35(1) of UK GDPR states that organisations must complete a DPIA where a type of processing is likely to result in a high risk to the rights and freedoms of individuals. For large-scale data processing activities, especially those involving health data, this may often arise. Where the released synthetic dataset is anonymous in the relevant hands, its release does not itself require a DPIA under UK GDPR. However, organisations should separately assess whether any earlier personal data processing to generate synthetic data requires a DPIA. If the release of the synthetic dataset does not require a DPIA, organisations should still document their risk assessment and decision-making as to how the conclusion was reached that the data was not personal.

Can I post my synthetic data openly on the internet?

Where synthetic data has been assessed as anonymous in the context of open release and appropriate governance checks have been completed, organisations may choose to release it openly, subject to a careful assessment of residual risk and appropriate governance review.

How do I need to label my synthetic data when I release it?

Synthetic data should be clearly labelled as synthetic in its title as well as in each variable heading (for example, adding 'synth' as a prefix or suffix). Both acceptable and prohibited uses of the synthetic data should be explicit. Where synthetic data is low-fidelity and not intended for decision-making or generating conclusions, this should be clearly stated. There is published guidance on how to label metadata and variable names when releasing synthetic data (Raab et al, 2025).

Is the advice the same for commercial and public sector organisations?

Much of the advice is the same for all types of organisations. The biggest differentiator is in the legal basis used to create the synthetic data. For public sector organisations, this may often fall within 'public task'. For commercial organisations, they may require, for example, 'legitimate interests' or 'consent' as legal bases.

Can I still release my data openly if I developed it with a machine learning model?

Possibly, but only where the released dataset has been assessed as anonymous in the context of open release. Many off-the-shelf models and workflows are designed for utility, not to protect privacy. Because these models are often hard to explain, considering the synthetic data generation method alone is often insufficient to demonstrate anonymity. There is no consensus on what metrics should be used to evaluate disclosure risk and there are new risk measures emerging. Organisations should use both quantitative metrics and qualitative evaluations to check disclosure risk. Organisations can decide that their synthetic data is anonymous and suitable for open release provided that they have residual risks and governance checks in place, and they follow good practice such as by minimising identifiable data entering the model, considering limiting the exposure of outliers within the dataset and comparing a range of disclosure risk metrics. Organisations should consider a 'motivated intruder test' in the context of the specific release setting.

I see synthetic data generation companies claiming that their data is out of scope of UK GDPR - how can I trust that?

Claims that synthetic data is out of scope of UK GDPR may be true, but should be assessed carefully on a dataset-by-dataset basis. The two main things to examine are the method used to generate the synthetic data and the context-specific evidence used to assess disclosure risk in the intended release setting. While some companies advertise prominent "privacy scores", it should be emphasised that there is currently no consensus on the best metric, and no single metric should be relied upon in isolation. Whilst quantitative metrics can be a useful indicator of risk, organisations should only consider these in addition to tests such as the 'motivated intruder test' and means reasonably likely to be used. If a synthetic data company is claiming anonymity, this needs to be tested in a given context with evidence of such testing. These tests may include carrying out feasible attacks on the data to determine if people could be identified from it. Organisations should therefore ensure they have confidence in the vendor's methods and evaluation approach, and that these align with the principles and governance measures outlined in this document. Just because data is out of scope of UK GDPR does not permit unlimited sharing or use of the data as other contractual and legal instruments may apply.

Do organisations need to inform people that their data is being used to create synthetic data?

Using personal data to generate synthetic data is a form of processing and is subject to the transparency requirements in Articles 13 and 14 of UK GDPR. However, synthetic data generation itself does not create additional notification obligations beyond those requirements. ICO guidance on transparency when publishing anonymised data should be consulted (ICO, What accountability and governance measures do we need? 2025).

Who decides whether synthetic data is personal data or not?

This decision rests with the data controller or joint controllers, taking into account the specific context in which the data will be used and shared. While external vendors may provide technical assessments or supporting analysis, responsibility for the final decision cannot be delegated.

Can anonymous synthetic data still be subject to confidentiality or contractual restrictions?

Yes. Even where synthetic data is anonymous, obligations relating to confidentiality, intellectual property, commercial sensitivity, or ethical commitments may still apply.

If synthetic data is considered personal data and still subject to UK GDPR, then does it really make data sharing easier?

It depends on the circumstances. For example, using synthetic data as a privacy-enhancing technology can help organisations meet the balancing test for data sharing under Article 6(1)(f). It can also help organisations support the safeguards expected under Article 89(1) for research-related processing and in that way, help to open up access to data that may previously have been completely inaccessible for research. Using synthetic data may support the application of the Data Minimisation Principle and Data Protection by Design (Article 25) but if synthetic data is considered personal data, a lawful basis and other applicable compliance steps are still required.

What happens if disclosure risk is discovered after release?

Disclosure risks will evolve as analytical techniques and external data sources change. It is expected that organisations continually review the disclosure risk of their synthetic datasets accordingly. Where an organisation has taken reasonable and proportionate steps to assess and mitigate risk at the time of release and thereafter, the emergence of new disclosure risks does not automatically imply a failure of compliance. Organisations should respond by reassessing the dataset, updating safeguards as appropriate, reviewing their governance processes and considering regulatory body notification where appropriate. ICO anonymisation guidance should also be consulted: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/what-accountability-and-governance-measures-do-we-need/#whattypeofdisclosure>

What are the essential requirements for releasing synthetic data?

At a minimum, organisations should be able to show that they have defined the intended purpose, users, release model and fidelity level; assessed the lawfulness of any upstream personal data processing; considered whether a DPIA is required; evaluated identifiability in the intended release context; made a documented determination of the dataset's legal status under UK GDPR in that context; applied proportionate governance, access and contractual controls; labelled the data clearly as synthetic with appropriate use limitations; kept adequate records and an audit trail; and put in place post-release monitoring and withdrawal or review processes in case risks change over time.

Glossary of Terms

Many of the terms below are context dependent and should be interpreted in line with UK GDPR, ICO anonymisation guidance and the intended release setting described in this framework.

Anonymous synthetic data

Synthetic data that, in the relevant hands and release context, does not enable individuals to be identified taking account of singling out, linkage and inference and the means reasonably likely to be used.

Article 9 condition

The condition under UK GDPR that must be identified where special category data are processed.

Controlled release

A release model in which access to synthetic data is subject to strict access controls and governance measures.

Data controller

The natural person or legal entity responsible for deciding the purposes and means of processing personal data. Responsibility for assessing whether synthetic data constitute personal data under UK GDPR rests with the controller or joint controllers.

Data minimisation

The principle that personal data used in generating or evaluating synthetic data should be limited to what is necessary for the stated purpose.

Data Protection Impact Assessment (DPIA)

A process required under Article 35 UK GDPR where a type of processing is likely to result in a high risk to the rights and freedoms of individuals. DPIAs are a key accountability mechanism and may be required for personal data processing involved in generating synthetic data, even where the released synthetic output is later assessed as anonymous.

Disclosure risk

The risk that something could be learnt about one or more individuals from data that is supposed to be anonymous. This could involve identification (understood to be singling out), linkage to other data and or attribution (where attributes are inferred from that data).

External datasets

Datasets outside the released synthetic dataset that may be available to recipients or adversaries and may increase disclosure risk if combined with the synthetic data.

Explainability

The extent to which the operation, assumptions and outputs of a synthetic data generation process can be understood and meaningfully described by those responsible for its use and governance. Explainability is relevant to assessing risks, accountability and appropriate safeguards, but does not by itself determine whether synthetic data are anonymous under UK GDPR.

Fidelity

The degree of closeness between the source data and the synthetic data. In this document, fidelity is relevant to assessing how far the synthetic data reflects the source data, but it is not the same as disclosure risk and does not by itself determine whether the data is anonymous.

Inference

The ability to use a set of attribute values from an individual in the source data, to infer, with high confidence, sensitive unknown information about the same individual using the synthetic data.

Lawful basis

A legal justification under Article 6 of UK GDPR that permits the processing of personal data. Where personal data are used to generate synthetic data, a lawful basis must apply to that upstream processing. If the generated synthetic data is considered personal data, a lawful basis may also be required for further processing.

Legitimate interests

A lawful basis under Article 6(1)(f) UK GDPR that applies where processing is necessary for the purposes of legitimate interests pursued by the controller or a third party, except where those interests are overridden by the rights and freedoms of individuals. Use of this basis requires a documented balancing assessment.

Linkability

The ability to link at least two records concerning the same Individual. In practice, this involves an adversary connecting a record in the synthetic dataset to a separate, external dataset by matching shared characteristics. A successful link reveals new sensitive information from the synthetic data (like income or health status) that was not present in the other data.

Means reasonably likely to be used

The means that a person or organisation could realistically be expected to use to identify an individual, taking account of factors such as available technology, cost, time, expertise, motivation and foreseeable technological developments. This assessment is context dependent and must consider who will receive or hold the data.

Motivated intruder test

A context-based test used in anonymisation assessment to consider whether a person without specialist privileges, but with reasonable determination and available resources, could identify individuals.

Open release

A release model in which synthetic data is made openly available. This is a particularly risky mode of release and should be subject to an assessment that residual risk is acceptable and that appropriate governance checks have been completed.

Outlier records

Records that are unique, rare or otherwise unusual and may therefore require particular scrutiny in disclosure-risk assessment.

Personal data

Information relating to an identified or identifiable individual within the meaning of Article 4(1) UK GDPR.

Post-release monitoring

The ongoing review of a released synthetic dataset to identify new risks, misuse, or changes in the external environment that may affect disclosure risk or governance requirements.

Public task

A lawful basis under Article 6(1)(e) of UK GDPR that applies where processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.

Purpose creep

The use of synthetic data beyond the purpose or acceptable uses originally defined by the releasing organisation.

Re-identification

The process by which an individual is identified from data that is claimed or intended to be anonymous, whether through singling out, linkage with other data, inference of attributes, or a combination of these mechanisms.

Release model

The manner in which synthetic data is made available to others, including the intended recipients, access conditions and governance controls. The release model is a key contextual factor in assessing identifiability and determining whether the data constitutes personal data in the relevant hands.

Residual risk

The risk that individuals could be identified, inferred about or otherwise harmed after reasonable and proportionate technical and organisational measures have been applied. Under a risk-based approach, decisions about releasing synthetic data depend on whether residual risk is acceptable.

Safeguarded release

A release model in which access is wider than a tightly controlled setting but still subject to governance measures, conditions or restrictions.

Singling out

The ability to isolate a record, usually from unique combinations of attributes, that distinguish an individual from the other records.

Source data

The dataset which is used to train the model which generates the synthetic data. In data synthesis we wish to retain the statistical properties of the original data (utility) whilst protecting the confidentiality of contributors to that dataset.

Special category data

Personal data revealing particularly sensitive information about an individual, such as health data, biometric data or data concerning racial or ethnic origin, which receive additional protection under UK GDPR and require an Article 9 condition to be met.

Utility

The usefulness of synthetic data for its intended purpose.

“Whose hands?” analysis

The assessment of identifiability by reference to who will receive or hold the data, recognising that the same dataset may have a different legal status in different hands.

References

UK Synthetic Data Community Group. Perspectives & Recommendations on the Development of Synthetic Datasets in Trusted Research Environments. 2026. <https://portal.dementiasplatform.uk/wp-content/uploads/2026/01/VSTAR-Framework-Report-Final.pdf>

Information Commissioner's Office. Anonymisation guidance. 2025. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/>

Information Commissioner's Office. What accountability and governance measures do we need. 2025. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/what-accountability-and-governance-measures-do-we-need#ensuretransparency>

Information Commissioner's Office. Examples of processing 'likely to result in high risk'. 2023. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/data-protection-impact-assessments-dpias/examples-of-processing-likely-to-result-in-high-risk/>

Information Commissioner's Office. How do we ensure anonymisation is effective? 2025. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/anonymisation/how-do-we-ensure-anonymisation-is-effective/#motivatedintruder>

Information Commissioner's Office. What is personal data? <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/>

Information Commissioner's Office. The Data (Use and Access) Act 2025: what does it mean for organisations? <https://ico.org.uk/about-the-ico/what-we-do/legislation-we-cover/data-use-and-access-act-2025/the-data-use-and-access-act-2025-what-does-it-mean-for-organisations/>

UK Government. Guidelines and best practices for making government datasets ready for AI. <https://www.gov.uk/government/publications/making-government-datasets-ready-for-ai/guidelines-and-best-practices-for-making-government-datasets-ready-for-ai>

Raab, G. M., McCall, S., and Cavin, L. Four checks for low-fidelity synthetic data: recommendations for disclosure control and quality evaluation. International Journal of Population Data Science, 10(2), 2025. <https://ijpds.org/article/view/2972>

SAFEHR. Synthetic Data Policy. <https://www.safehr-data.org/synthetic-data-policy>

DPUK Data Portal / UK Synthetic Data Community Group. Perspectives & Recommendations on the Development of Synthetic Datasets in Trusted Research Environments. <https://portal.dementiasplatform.uk/reports/development-of-synthetic-datasets-in-trusted-research-environments/>

Hotchkiss, L., Adeoye, K., Squires, E. and Thompson, S. (2025). Developing Synthetic Data Tools for Trusted Research Environments to Enable Researcher Training. International Journal of Population Data Science, 10(4). <https://ijpds.org/article/view/3096>

ADR UK Synthetic Data Working Group. Considerations for the provision of synthetic forms of secure data: An ADR UK Synthetic Data Working Group Guidance paper. Published December 18, 2025. <https://zenodo.org/records/17980237>

Magder, C., Haaker, M., Kasmire, J., Zahid, H. and Ogwayo, M. (2025). The Role of Synthetic Data in Research: Benefits, Costs, and Practical Insights from Data Owners and Trusted Research Environments Experts – A Project Report with Practical Recommendations. <https://zenodo.org/records/15191271>