Research Paper: Exploring Synthetic Data Validation – Privacy, Utility and Fidelity

Disclaimer - The below insights reflect the views and discussions of participants of the roundtable, and should not be taken as the official view of the Financial Conduct Authority, the Information Commissioner's Office, or the Alan Turing Institute.

Contributors:

Valerie Marshall (Financial Conduct Authority)

Charlie Markham (Financial Conduct Authority)

Pavle Avramovic (Financial Conduct Authority)

Paul Comerford (Information Commissioner's Office)

Carsten Maple (The Alan Turing Institute)

Lukasz Szpruch (The Alan Turing Institute)

Contents

Summary 1

Background 2

Overview of the day 3

Key insights: 4

1. **Validating utility and fidelity** 4
2. **Validating privacy** 6
3. **Other approaches to advancing synthetic data** 8

Next steps & Conclusion 9

Appendix 10

Summary

- The below paper explores the insights from a roundtable event hosted by the Financial Conduct Authority, the Information Commissioner's Office, and the Alan Turing Institute.
- The challenge of validating synthetic data from utility, fidelity and privacy perspectives remains a critical barrier to synthetic data adoption.
- Understanding the requirements of a specific use case is paramount when assessing both utility and privacy.
- Model generalisability may be one method to increase the utility of synthetic data generators across multiple use cases, however this can lead to challenges with model drift and re-identification.
- Mathematical methods that validate the model that generated the synthetic dataset (the synthetic data generator) need to be supplemented with post-generation validation of the synthetic dataset.
- To drive adoption, the industry could shift to a risk-based model for privacy validation that accepts some level of inherent risk in generating and sharing synthetic data.

- Beyond validation, there are several ways to advance synthetic data adoption including use case documentation, standards and frameworks for adoption, and regulatory guidance.

Background

In an era of increasingly digitalised financial services, the financial services industry generates a vast amount of data. This data has the potential to drive innovation in financial services while also improving the efficiency and effectiveness of products and services in the sector. However, the need to protect peoples' privacy places conditions on sharing this data.

In response to these challenges, regulators, industry, and research organisations have explored the potential of synthetic data (and other privacy-enhancing technologies (PETs)) to enable the sharing of data to drive safe and responsible innovation in financial services.

Synthetic data uses a mathematical model or algorithm to generate statistically realistic, but 'artificial' data. Not only does synthetic data enable data sharing in a privacy-preserving manner; it can help organisations to better harness the power of their data by mitigating data quality issues, modelling new and emerging scenarios, and protecting commercially sensitive data.

The Financial Conduct Authority (FCA), the Information Commissioner's Office (ICO) and the Alan Turing Institute (Turing) each have active work programmes on synthetic data:

The Turing is working with a range of partners, including HSBC, UCLH, Accenture and the Office for National Statistics, to develop novel techniques for the generation of synthetic data in a wide range of use cases. The Turing is committed to the development of tools, techniques and policy for the effective generation of synthetic data. Representatives from the Turing sat on the Royal Society working group on Privacy Enhancing Technologies, and members of the institute authored the Royal Society-commissioned paper *Synthetic Data: What, Why and How?* which identified the landscape of uses, techniques and measures for synthetic data generation.

The ICO is encouraging the use of PETs, as these technologies can unlock the economic and societal benefits of data sharing while incorporating a data protection design and default approach to the use of data. Sharing financial data can carry significant risk of harm to individuals, both because transaction and spending patterns can reveal information about an individual's private life, and because of the financial harm that can be caused from unauthorised access to banking information. Synthetic data is an effective technique to facilitate safer use of financial data by preserving patterns in the data, while mitigating the risk of any individual being identified from it. The ICO have produced guidance on privacy enhancing technologies which provides further information on how the use synthetic data supports compliance with data protection law and considerations for implementation. Additionally, we are working with organisations to develop purpose-specific case studies, illustrating how synthetic data can be used in practice.

In March 2022, the Financial Conduct Authority (FCA) expanded its engagement with synthetic data by publishing a Call for Input to gather views from industry and academia on the potential of synthetic data to expand data access and drive innovation in financial services. While the feedback emphasised the potential of synthetic data, it also highlighted some significant challenges to the adoption of the technology in the financial services sector. Among these challenges, respondents emphasised the difficulty of validating synthetic data from both a privacy and utility perspective as a critical barrier

to adoption in financial services. Furthermore, responses to the Call for Input revealed that there is limited consensus in both academia and industry on the best techniques for validating synthetic data in different contexts, and a lack of common language for discussing validation techniques.

To address these challenges and advance responsible innovation, the FCA, the Turing and the ICO partnered in March 2023 to host a joint industry and academic roundtable on validating synthetic data. This event convened experts to discuss the challenges of validating synthetic data and to progress early thinking towards potential solutions.

Participants represented a broad spectrum of organisations, including financial institutions, synthetic data vendors, regulatory agencies, and academic institutions. Each contributed their unique perspectives and expertise to the conversation.

This joint paper provides an overview of the roundtable, capturing key insights from the event, including validation from utility and privacy perspectives, as well as additional ideas on overcoming barriers to the adoption of synthetic data in financial services. The paper concludes with each organisations' next steps for their synthetic data work programme and suggests avenues to foster ongoing collaboration to address barriers to synthetic data adoption in the financial services sector.

By bringing together the expertise of the FCA, the Turing, and the ICO, this paper aims to contribute to the ongoing research and dialogue on synthetic data to support responsible innovation in financial services.

Overview of the day

The roundtable had various objectives, from gathering a diverse community to discuss challenges associated with synthetic data validation, to highlighting broader barriers to synthetic data adoption in industry. Validating synthetic data – ensuring it remains useful for analysis and decision-making whilst also maintaining the necessary levels of privacy over time – is a complex challenge for synthetic data generation. This process may require different techniques and considerations depending on the use case, the acceptable level of risk and the accuracy requirements.

To understand the various considerations for this challenge, the event incorporated a mix of panel discussions and breakout sessions. The roundtable discussions centred on the following questions:

- How can organisations evaluate **fidelity** and **utility** in synthetic datasets, and what criteria should be used to determine a "good" synthetic dataset?

- What are the most effective **techniques** for **validating privacy** in synthetic datasets, and how can organisations strike a balance between privacy protection and data utility?

- What broader challenges to the adoption of synthetic data must be addressed, and what potential solutions can be proposed?

Definitions

Utility: a synthetic dataset's 'usefulness' for a given task or set of tasks, for example for training AI or Machine Learning models.

Fidelity: refers to measures that directly compare the synthetic dataset with the real dataset i.e. the statistical similarity of the synthetic dataset to the input real data.

Privacy: measures the risk that specific individuals (or other sensitive data) can be re-identified from the synthetic dataset.

Key insights:

1. Validating utility and fidelity

The use case is central to discussions around utility and fidelity

Participants of the roundtable emphasised that assessing both utility and fidelity will depend on the purpose of the synthetic data. Defining a clear and agreed set of goals prior to the generation process will benefit further discussions on utility and fidelity (noting that fidelity is not always required for a synthetic dataset to have utility).

The need to generate and assess synthetic data for each use case could limit incentives to test and adopt this technology. Participants noted that model generalisation – the ability of a deep learning model to learn and properly predict the pattern of unseen data – could be important to drive the adoption of synthetic data. Generalisation could enable models that generate synthetic data (synthetic data generators) to synthesise data outside of their initial training datasets. Participants felt this could be a potential avenue to expand the utility of synthetic data generators beyond a single use case.

However, challenges such as model drift – the degradation of model performance due to changes in the underlying distribution of data, the target variable, and the relationship between input and output variables – could limit the use of synthetic data to a case-by-case basis. Generalisation may also reduce the privacy of a synthetic dataset, which we discuss below (see "The importance of the use case when validating privacy").

Whether synthetic data generators can be generalised to meet multiple use cases is a growing area of industry and academic debate.

Approaches to validating synthetic data

Discussions around validating utility often differentiate between 'broad' and 'narrow' measures. 'Broad' measures assess the fidelity of the synthetic dataset to the real dataset by quantifying the statistical similarities between the training data and the synthetic dataset. 'Narrow' measures compare the differences in model performance, for example inference or prediction, between the original and synthetic dataset.

Broad measures

Participants highlighted that using 'broad' measures to retain the statistical properties of the original dataset could enable the use of synthetic data for multiple use cases that require a double of the original data. Measures to achieve statistical similarity may include comparing univariate and multivariate distributions and the correlations between features.

However, the more similar the synthetic data is to the real data, the higher the risk of re-identification, potentially to the extent where it becomes too difficult to use or share the synthetic dataset. This outcome would drastically reduce the utility of the synthetic dataset.

Narrow measures

'Narrow' approaches premise that there is no need for a synthetic dataset to perfectly replicate the real data to have value. Instead, narrow approaches assess the utility of synthetic data by comparing the results of a model trained on a synthetic versus real dataset. However, assessing whether the output of a model trained on synthetic data is sufficiently similar to a model trained on real data can be difficult without an industry standard or a detailed understanding of the acceptable range of difference in model performance.

Mathematical approaches

Participants outlined an additional theoretical basis for validating synthetic data from both a utility and fidelity perspective. They discussed whether it is possible to build mathematical guarantees of fidelity into the synthetic data generator itself, as opposed to a post-hoc assessment of the synthetic dataset or a model trained on synthetic data (as with broad and narrow measures).

Additional explanation

Building fidelity and utility into the synthetic data generator is an area of active research. Part of this research focuses on the role of loss functions in the generation process, which measure how far an estimate value is from its true value. For synthetic data, loss functions could measure the degree of difference between the synthetic dataset and the input data.

In the case of fidelity, developers may base loss functions on a combination of univariate and multivariate measures of fidelity, which the synthetic data generator will then try to minimise. This approach may increase the chances of a high-fidelity output by building mathematical guarantees of fidelity into the model.

Whilst this approach could work well for straightforward loss functions, models with complex loss functions (that capture many measures) may struggle to reliably converge to a solution. Such models may require additional computing power to confidently assess that their solution is optimal.

The challenge with this approach would be convincing legal, compliance and/or senior stakeholders that these mathematical 'guarantees' are reflected in practice in the generated dataset, without confirming these results with more tried and tested post-generation assessments. It is likely that post-hoc validation methods (broad or narrow approaches) would still be required to build trust in the synthetic dataset.

Qualitative approaches

Qualitative approaches to synthetic data validation may be used to supplement the techniques outlined above. Participants acknowledged that techniques to validate the utility and fidelity of synthetic data are still nascent. Developers may therefore need to draw on additional expertise to qualitatively assess whether the data is accurate.

For a fraud use case, for example, fraud or financial crime experts may be required to assess whether the fraud typologies present in the synthetic dataset match known

typologies in the industry. Typologies are a way of describing groups that display different clusters of behaviours or attitudes.

To achieve this, developers would need to standardise the typology formation and explain to experts how the data is behaving in relation to the typology. Bridging this gap and speaking in the same language could represent a challenge for technical and subject matter experts.

Whilst this approach should be considered as an addition (not replacement) to the approaches outlined above, accessing use case-specific expertise could represent an additional challenge to synthetic data generation.

2. Validating privacy

The importance of the use case when validating privacy

Validating the privacy of synthetic data should begin by assessing the use case and the required level of accessibility. For example, the use of synthetic data for internal research and development may (although not always) require less stringent privacy guarantees than synthetic data that is being shared externally. In such cases, utility and/or fidelity may have greater consideration during the generation process. Organisations should consider how different forms of synthetic data can pose different identifiability risks, and choose an appropriate release model to mitigate them.

In addition, organisations may differentiate between use cases that require the processing of *confidential* data versus *personal* data:

- **Confidential information** is defined by each individual organisation and may include non-public, sensitive and/or business-related information
- **Personal data** is information that relates to an identified or identifiable individual and is subject to specific requirements under UK GDPR.

Organisations in breach of personal data requirements risk regulatory fines, customer compensation claims, legal fees and reputational damage. They may therefore adopt a higher privacy risk threshold when processing personal data to generate synthetic data.

To reduce the risk of re-identification from a synthetic dataset, and align with the UK GDPR's principles of data minimisation and purpose limitation, organisations should only include the properties needed to meet their specific use case, and nothing else. Organisations need to assess whether their synthetic dataset is anonymous or identifiable. This approach should be prioritised over adding additional properties to enhance the generalisability of a synthetic dataset.

To navigate the trade-off between privacy and fidelity, organisations might group potential use cases in terms of the type of fidelity and features required and generate multiple synthetic datasets, defining specific privacy requirements for each dataset. In line with UK GDPR, organisations that take this approach will need to ensure that, where they use personal data to meet additional use cases, this is compatible with the original purpose, they have the consent of the data subjects, or they have a clear obligation or function set out in law.

Validating the privacy of the data versus the generator

Participants debated the various benefits and trade-offs between mathematical notions of privacy (such as differential privacy) versus post-generation assessment of the synthetic dataset. The debate presided over whether it is possible to build privacy

guarantees into the synthetic data generator, or whether post-generation testing of the synthetic data is always required. One participant mentioned that post-hoc testing may take a similar form to penetration testing, whereby an organisation's cyber team attempt to re-identify individuals present in the original dataset.

Mathematical evaluations are useful tools to detect possible flaws in an algorithm or its implementation, but may lead to false guarantees of privacy when there is none. The 'Infinite Monkey Theorem' (more formally called the Law of Large Numbers), for example, states that a monkey hitting keys at random on a typewriter for an infinite amount of time will eventually replicate any given text, such as the complete works of William Shakespeare. Applying this theory to synthetic data, one could deduce that a synthetic data generator will eventually generate an exact replica of the real data by chance. Whilst the likelihood of this occurring is low, it does mean that post-generation testing of the synthetic dataset is needed to affirm mathematical notions of privacy.

Participants noted that convincing legal teams and senior stakeholders that the generated data is actually private is an additional challenge with mathematical notions of privacy. One participant used crash testing as an analogy to illustrate this point. They stated that customers would feel much safer driving a car that has been thoroughly crash tested over a car that had mathematical 'guarantees' of safety built in when it was being developed. It follows that legal and senior stakeholders will likely prefer additional post-generation testing to confirm that the synthetic dataset is sufficiently private.

Transitioning to a risk-based model for privacy

Regulators and standard-setting bodies are often asked about precise numbers, metrics or criteria that can guarantee the privacy of a synthetic dataset.

In reality, the only complete privacy guarantee that can be given is if organisations do not create or share synthetic data. As a rule of thumb however, zero risk equates to zero utility for a synthetic dataset.

To drive adoption, the industry could shift to a risk-based model that accepts some level of inherent risk in generating and sharing synthetic data. This approach aligns with the UK GDPR, as data protection law does not require anonymisation to be completely risk-free. Organisations must be able to mitigate the risk of re-identification until it is sufficiently remote that the information is 'effectively anonymised'. They must assess whether the identification of individuals is 'reasonably likely' relative to the circumstances of the processing, rather than protect against every hypothetical or theoretical risk of identifiability.

A risk-based model may include risk frameworks and acceptable ranges of risks. Factors assessed in such a model may include singling out, linkability and inference, the release model (public or closed release), motivated intruders, and the time, cost, and technologies/techniques available for re-identification. These ranges may be defined by an independent body, such as the IEEE Synthetic Data project, other industry standard-setting initiatives, and/or regulators. Industry might also need to socialise risk frameworks and ranges with stakeholders within their organisation (seniors, legal and cyber teams) to build trust in the overall approach.

Risk is not static, therefore, a risk-based model for assessing the privacy of synthetic data will need to be iterative. For example, the risk of re-identification for a particular synthetic dataset may increase over time as more real data becomes publicly available. The feasibility and cost-effectiveness of the available means to identify individuals can also change over time. This means the status of synthetic data – as personal data or anonymous – can change over time.

Moving to and gaining widespread stakeholder buy-in for a risk-based model could help overcome barriers to adoption for synthetic data and build trust in the technology across the data lifecycle.

3. Approaches to advancing synthetic data

This section explores broader approaches for encouraging the adoption of synthetic data in financial services. It investigates three interdependent approaches discussed during the roundtables: Use Case Exploration; Standards, Frameworks and Guidance; and the Role of the Regulator.

While these approaches are separated in this paper, we recognise their interdependency. Advancing these three approaches can help build trust in this emerging technology, ensure responsible use in financial services, and drive further synthetic data adoption in the sector.

Use Case Documentation:

The most frequently mentioned approach during the roundtables was the exploration of use cases to expand the evidence base for the adoption of synthetic data. Ultimately, participants agreed that synthetic data is not a silver bullet for sharing data in financial services. Given this, deeper investigation into potential use cases of synthetic data is required to drive adoption in the financial services sector and determine when it's relevant, useful, and when it may not be suitable.

Participants also acknowledged the potential for collaboration between industry, academia, and regulators in producing more tried and tested use cases for synthetic data. Participants referenced anonymisation techniques, testing the validity of synthetic datasets for specific purposes (e.g., identifying fraudulent activity), and augmenting existing datasets to improve their utility for certain financial services applications.

Furthermore, participants noted that effective use cases require thorough documentation. As a minimum, such documentation should include details on how utility was measured, fidelity considerations, privacy trade-offs, and potential biases of a model. This documentation can fill knowledge gaps, drive synthetic data development, and provide stakeholders with essential information to determine if synthetic data is the right tool for their use case.

Participants agreed that developing and sharing use cases creates a larger evidence base for stakeholders, building trust in this emerging technology and increasing the buy-in needed for synthetic data adoption in financial services. To this end, the FCA will continue exploring ways to facilitate use case exploration, potentially leveraging its Innovation Services (such as TechSprints or the Digital Sandbox) and convening experts to work collaboratively on targeted, non-competitive use case development.

The ICO is committed to facilitating the development of the responsible and legal use of synthetic data use cases in financial services. Working with external experts from industry and academia, we are developing case studies which will demonstrate how synthetic data can be used to share data responsibly and legally between financial institutions, while complying with data protection law and mitigating the risks to individuals.

Standards, Frameworks and Guidance:

The IEEE outlines that standards form the fundamental building blocks for product development by establishing consistent protocols that can be universally understood and adopted, while also helping to build consumer trust and stakeholder buy-in.

However, attaining industry standards first requires collaboration among stakeholders to develop a comprehensive understanding of how firms are using synthetic data in industry. Throughout the roundtable, we observed a high level of interest from industry and academia to engage and to advance discussions to inform future conversations on standards in synthetic data.

Building upon the further exploration of use cases and potential standardisation, participants highlighted the need for practical frameworks and guidance to support synthetic data vendors and industry players in financial services. These resources would not only facilitate a better understanding of validating synthetic data from the perspectives of utility, fidelity, and privacy but also provide practical, step-by-step guidance for experimenting with this technology.

Frameworks and guidance can equip practitioners with the confidence to navigate the application of synthetic data, ensuring the responsible use of emerging technologies and techniques. Moreover, these resources could help drive stakeholder buy-in by providing evidence for the technology's effectiveness, thus fostering responsible adoption within financial services.

Role of the Regulator:

The Regulatory Horizons Council 'Closing the Gap' report highlights how regulation and regulators play an important role in a landscape that supports innovation in emerging technologies, such as synthetic data. Regulators can serve as conveners of stakeholders to create non-competitive environments that drive collaboration and innovation in synthetic data. For example, participants mentioned the benefit of the FCA's Innovation Services, such as TechSprints and the Digital Sandbox, for driving innovation in synthetic data.

Participants recognised that, while further guidance from regulators is essential as a technology matures, it is equally important for regulators to provide space for the creation of industry-led guidance, frameworks, and standards. In light of this, the ICO and Professor Carsen Maple of the Alan Turing Institute are participating in the IEEE Synthetic Data standards working group, to provide a regulatory and academic perspective to the industry effort. Regulators can also bring the ecosystem together to collaborate on the unknowns, barriers, and challenges associated with synthetic data in financial services.

m

Next steps & Conclusion

This paper has discussed the validation of synthetic data from a utility, fidelity and privacy perspective, including the importance of the use case to synthetic data validation, the different approaches to assessing utility and fidelity, and the value of shifting to a risk-based model for validating privacy.

Alongside the insights discussed above, the roundtable highlighted several outstanding challenges and remaining questions, not least around how to continue building stakeholder trust in the technology. Participants also discussed the role of bias during the generation process, and the ethics of 'debiasing' data to reflect the world we might like to see versus finetuning the generator to account for bias. Participants also noted that more granular discussions into specific use cases, focusing on the associated challenges and the necessary steps to overcome these challenges, could be a fruitful route forward.

These remaining questions provide a stimulus for future reflection on synthetic data. We will further explore these themes through the FCA's recently established Synthetic Data Expert Group, which comprises 22 members across financial services firms, academia, synthetic data vendors, public sector organisations and others. The group will develop best practice around a set of specific synthetic data use cases to provide practical guidance to firms and innovators looking to test this technology. The group will also develop a framework for collaboration between industry, academia, public sector and other relevant stakeholders to drive further exploration and testing of use cases in a collective manner.

The ICO is currently working with the IEEE synthetic data standards group to develop a working paper on best practices for synthetic data generation, including legal considerations when generating synthetic data. Once complete, the IEEE group plan to develop this into the first complete standard for synthetic data generation. The ICO will continue to engage with key stakeholders, including international counterparts, and look for further opportunities to develop relevant synthetic data case studies in the financial services sector. The ICO is also a member of the FCA's Synthetic Data Expert Group, within which our aim is to ensure that data protection considerations are integrated into the work of the group.

The Turing is developing techniques for the synthetic generation of a range of data type including time-series, graphs, tabular and biometrics including faces. It is developing novel pipelines, including a privacy risk-based approach to increasing fidelity, for the development of data. It is also developing techniques to measure the efficacy, privacy and fidelity of synthetic data. To ensure the relevance and outreach of its work, members of the Turing will continue to contribute to the FCA's Synthetic Data Expert Group and on the IEEE synthetic data standards group.

According to Gartner, synthetic data is at a critical peak of development, and the next one to two years will be decisive to the future progress of this technology. We look forward to future engagement and collaboration between industry, academia and the public sector to advance this technology from a point of theory to practice.

Appendix

List of attending organisations

BetterData

Financial Conduct Authority

FinCrime Dynamics

Hazy

HSBC

Imperial College London

Information Commissioner's Office

Lloyds Banking Group

Lucinity

MostlyAI

NatWest

NayaOne

Ofcom

Smart Data Foundry

Statice

SWIFT

The Alan Turing Institute

The Office for National Statistics

The Royal Society of Science

UK Finance

University of Cambridge

University of Cardiff

University College London