Financial Conduct Authority

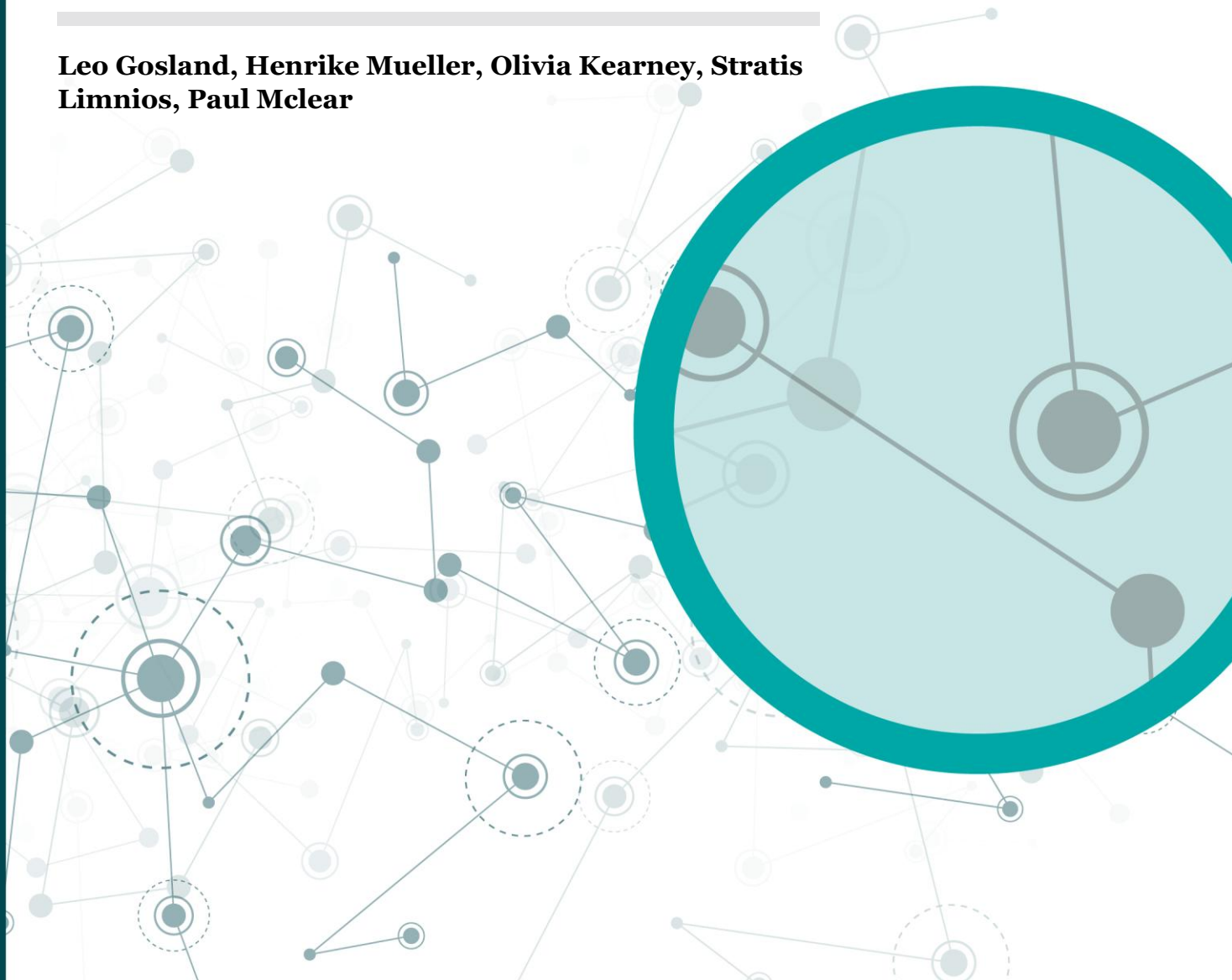
Research Note

15 April 2026

Synthetic Data and Anti-Money Laundering

Project Report

Leo Gosland, Henrike Mueller, Olivia Kearney, Stratis Limnios, Paul Mclear



FCA research notes in financial regulation

Disclaimer

Research notes contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. This note is provided for general information only. The FCA does not guarantee the accuracy, completeness, or reliability of this note. The FCA accepts no responsibility for any errors or omissions in this note, any loss or damage arising from reliance on this note, or for any action taken based on the information provided.

Authors

Leo Gosland and Henrike Mueller (FCA); Olivia Kearney, Stratis Limnios, Paul Mclear (Plenitude).

Acknowledgements

We would like to thank Paul Comerford (ICO) and Janet Bastiman (Napier AI) for their comments.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics@fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Contents

Summary

This report presents the findings of the FCA's Synthetic Data and Anti-Money Laundering project, which examines how synthetic data can support innovation in money laundering detection while safeguarding privacy.

The project aimed to create a statistically realistic, privacy-preserving dataset containing embedded money laundering typologies, which will enable firms to test and demonstrate new and emerging money laundering detection solutions in an upcoming Data Sprint without using real customer data.

The report concludes that well-designed synthetic data can provide meaningful analytical value, allowing firms and regulators to experiment safely with new and emerging detection approaches. It highlights important trade-offs between realism, privacy and coherence, and emphasises that synthetic data should complement, not replace, live operational data.

1 Overview

Purpose

Experts estimate that criminals launder between 2% and 5% of global GDP each year, equivalent to around USD 800 billion to USD 2 trillion. In response, financial institutions invest substantial time and resources in detecting and preventing money laundering.

To support innovation in this area, we initiated this project to develop a fully synthetic dataset containing embedded markers of money laundering. This dataset will be made available to participants in the upcoming Synthetic Data AML Solution Sprint, providing a safe environment in which to experiment and demonstrate new approaches.

We set two goals:

1. That the dataset should look and behave like real data, protect privacy, include realistic money-laundering typologies, and be useful for firms.
2. To demonstrate that new and emerging approaches to detect money laundering have value in the fight against financial crime.

We want more firms to experiment safely and innovatively to develop new solutions to detect money laundering. If they can, firms as well as regulators and government can more easily identify and disrupt illicit financial flows, reduce harm to society, and build public trust in the financial system.

We want the project to achieve the following:

- Vendors and fintech innovators can demonstrate that their AML tools work without needing access to sensitive real data.
- Regulators and policymakers can build evidence on how they can use synthetic data in supervision.
- The public and financial markets should benefit indirectly as better AML tools reduce the risk of illicit finance flowing through the UK economy, strengthening market integrity and public trust.

We hope this project shows that synthetic data can help regulators and firms work together, supporting beneficial innovation in UK financial markets.

This dataset is being made available through the FCA Digital Sandbox as part of the upcoming data sprint, to firms who are developing transaction monitoring solutions – with a particular focus on new and emerging techniques, such as AI.

For more details about the data sprint and how to apply, please visit the landing page here (https://events.fcainnovation.co.uk/Synthetic_AML_Solutions_Sprint), applications are open until 26 April.

2 Context

This project supports wider FCA priorities, including our Strategy 2025-30. We want to be a more data-led regulator, reduce and prevent financial crime, and promote innovation and competition across financial services.

It also supports the UK Economic Crime Plan and our wider work on AI, by building practical experimentation into supervision and policy work.

We used advanced technical methods to generate the dataset. This is some of the most ambitious data science work we've ever done, and it shows our ongoing investment in innovative new approaches.

Background

Tackling financial crime requires coordinated effort across multiple domains and the Synthetic Data and Anti Money Laundering project is a multi-stakeholder initiative bringing together the FCA, the Alan Turing Institute, Plenitude Consulting, and Napier AI.

Each partner contributes distinct expertise: we provide regulatory leadership, oversight and technical skills; the Alan Turing Institute brings research and technical experience in synthetic data; Plenitude contributes financial crime and industry expertise; and Napier AI provides applied technology and product experience in detection of financial crime.

The challenge of designing effective AML tools

Detecting and tracing suspicious behaviour

AML systems must be able to detect behaviour spread across multiple accounts, entities, and transaction types. A single suspicious transaction rarely tells the full story. Instead, it's the network of relationships and the layering of funds that can reveal illicit activity.

To test whether models can capture this, practitioners need datasets that mirror real-world financial transactions. Yet using real customer data can pose legal and ethical risks and it's particularly difficult to anonymise data fully without stripping away the patterns that may matter most.

Embedding money laundering typologies makes synthetic data useful for detecting and evaluating money laundering. Without being able to represent illicit behaviours, the data would offer little value.

Why use synthetic data?

Synthetic data can address some of the challenges above. Datasets that replicate the statistical properties of real financial transactions, while embedding realistic synthetic laundering typologies can support innovation without exposing sensitive information.

Synthetic data is valuable because it:

- protects privacy
- lowers barriers to experimentation
- can be shared easily
- acts as a more standardised basis for testing, training, and benchmarking
- reduces dependence on access to live banking data
- lets firms evaluate how their tools perform against known patterns of illicit behaviour

When combined with formal privacy guarantees, such as differential privacy, synthetic data can achieve the right balance between fidelity, utility, and privacy. It preserves the complexity of real behaviours, so that models can be trained and tested meaningfully, while ensuring that personal data is not compromised.

3 Methodology

Overview

We used the following methodology for generating the synthetic data sets:

1. We received real banking data which had been anonymised at source.
2. We augmented the data with synthetic money laundering typologies designed to reflect real world examples.
3. We used the anonymised source data including the encoded synthetic money laundering typologies to generate new fully synthetic datasets, using privacy-focused data science techniques described below.

Synthetic data and protecting privacy

We used the Adaptive and Iterative Mechanism (AIM) – a recognised approach for generating synthetic data[1].

It protects privacy by introducing controlled randomness to ensure that no individual customer or transaction can be reverse engineered from a dataset, while still preserving the patterns needed for meaningful analysis.

While synthetic data can raise residual privacy risks in principle (in relation to both individuals and firms), for this project those risks were fully managed and mitigated by:

- excluding personal details from the real data – no personal data was included in the initial data request
- applying privacy-preserving controls during generation

In addition to this, the synthetic dataset is only accessible during the data sprint to participating firms subject to strict controls and contractual obligations.

Typology generation

Including money laundering typologies – recognisable patterns of behaviour associated with illicit activity – adds value when combined with statistical realism, to ensure that the dataset resembles genuine banking activity.

Well-established typologies

Subject-matter experts defined recognised money laundering typologies, reflecting what compliance teams and regulators would expect to see in real cases.

These include behaviours such as:

- structuring transactions to stay below reporting thresholds

[1] **McKenna, Ryan, Brett Mullins, Daniel Sheldon, and Gerome Miklau. "AIM: An Adaptive and Iterative Mechanism for Differentially Private Synthetic Data."** *Proceedings of the VLDB Endowment (PVLDB)*, **15**(11), 2599–2612, 2022

- moving funds rapidly across multiple accounts to obscure their origin
- circular transaction patterns where funds return to their starting point via intermediaries.

Variations and complexity

To avoid creating overly simplistic or predictable patterns, the project also introduced variations around these typologies. This means that suspicious behaviour does not appear in identical forms each time but instead aims to reflect the diversity and complexity of real financial crime.

It's important that detection models are tested against realistic challenges rather than static rule patterns, so the dataset includes both well-established typologies familiar to practitioners and more complex variations that are harder to detect.

4 Testing and Evaluation

Once the data had been created, we completed tests against the data to assess whether the datasets were fit for purpose: does it preserve statistical fidelity, protect privacy, embed typologies effectively, and provide real-world utility for firms?

Statistical tests

Results showed that the synthetic data is a reliable substitute for its real-world counterpart. When we compared statistics of the anonymised source data with those of the synthetic dataset, there was little divergence, suggesting that the synthetic data reflected the statistical properties of the real data.

Importantly, not all observed behaviours could be cleanly attributed to a single cause. In some cases, it was difficult to disentangle whether patterns reflected the injected typologies, were artefacts of privacy-preserving design choices, or genuinely emergent behaviours.

This ambiguity reflects the complexity inherent in real financial data.

Privacy

Privacy protection was a core part of how the dataset was assessed. Throughout the data generation process, we ensured that the strength of privacy protection was sufficient. Stronger privacy protection can sometimes make the data less detailed, so oversight is needed to strike the right balance. This process provided a strong protection against re-identification of individuals, patterns and statistics present in the datasets.

Detecting money laundering typologies

Typology detectability is another important element of the evaluation. It’s not enough to include money laundering typologies in the datasets – firms must be able to detect them in realistic tests.

We tested whether the money laundering typologies could be detected within the data using established industry standard approaches. This helped us determine how easily they could be found – if they were too easy to detect, or too hard, the data sets would have limited value. The test results showed that there was a spectrum of detectability, which is the result we need for maximum utility in the data sprint.

The data sprint will test this further, and we'll use the results to improve how we encode typologies in future versions of the dataset.

The Data Sprint – real world evaluation

By making the datasets available as part of the data sprint and observing how firms interact with it, the project will show if the data can enable and demonstrate beneficial innovation.

It allows firms to demonstrate new money laundering detection approaches in a controlled but realistic environment, marking a shift away from isolated, siloed experimentation towards open and collaborative innovation.

The data sprint is an independent setting in which results can be observed and challenged, especially on whether the datasets reflect what practitioners have seen in real- world money laundering cases and demonstrating whether novel ways to detect money laundering are effective.

We'll use the feedback from participants to update and improve the datasets for future use cases.

5 Limitations and challenges

Balancing realism with privacy

One of the most persistent challenges of this project was how to represent real-world financial activity while ensuring that no sensitive information was exposed.

An example of this was in location data: postcodes carry high re-identification risk if reproduced in full. The team addressed this by replacing them with outcodes (the first segment only), retaining any data that appeared imperfect or counterintuitive.

Issues with data quality meant that addressing every anomaly risked producing datasets that were cleaner than reality and therefore less useful for testing AML systems. Where we couldn't resolve issues without undermining privacy or realism, we documented them rather than correcting them.

Ensuring internal coherence

Maintaining internal consistency across data elements was also challenging.

Money laundering behaviours often depend on relationships – between customers, their accounts, and their transactions. If those relationships are missing or incoherent, the embedded typologies become meaningless.

So the team created linked datasets in which profiles, accounts and transactions are generated as part of a coherent whole rather than as isolated rows. But certain inconsistencies remained, such as unusually high numbers of accounts associated with individual customers, inconsistent branch identifiers, and currency-related anomalies.

These data-level constraints sit alongside inherent modelling limitations. AIM generates transactions independently, which makes it hard for it to reproduce behaviours that depend on *sequences* of activity over time.

Rather than artificially inserting obvious transaction sequences, we selected synthetic accounts from the generated data and used them as the basis for embedding the typologies as naturally as possible in the synthetic dataset.

Caveating typologies

The datasets can only incorporate known typologies and inevitably, there are 'unknown unknowns': laundering techniques that criminals are already using but which have not yet been identified or codified. This means the datasets cannot capture the full spectrum of financial crime.

There are two lessons here: first, that synthetic datasets should be regularly updated as new typologies emerge, and second, that users should not treat the dataset as exhaustive, but as a training and testing tool to use alongside other datasets and analytics.

6 Risks – a forward look

Evolution of criminal behaviour

Financial crime is not static and the behaviours in these datasets reflect what we understand currently, rather than the full range of tactics criminals may already be experimenting with.

Typologies in the synthetic datasets are based on known behaviours – structuring, layering, roundtripping, high-risk jurisdiction transfers, and so on. But financial criminals continuously adapt, developing new techniques in response to technological change, regulatory pressure and enforcement actions.

The datasets could become outdated if they don't evolve alongside the threat landscape. So to remain relevant, they will require refreshes, incorporating emerging typologies and intelligence feeds.

Without this, there's a danger that firms may optimise their systems for yesterday's risks rather than tomorrow's.

Unintended consequences

Synthetic datasets can exhibit emergent properties – patterns that arise from the interaction of privacy processing, typology injection, and modelling choices, rather than being explicitly designed.

These introduce risks that must be carefully considered, because firms may respond to these artefacts as if they were genuine indicators of risk. This can distort testing, skew model development, and create false assurance about real-world performance.

Another potential risk is the possibility of firms optimising their systems to detect the specific typologies embedded in the dataset without improving broader detection capabilities.

A further unintended effect is misplaced confidence. If synthetic data is seen as a substitute for live operational data, there is a risk that firms could over-rely on it, neglecting the need for ongoing real-world calibration and validation.

It's important to treat synthetic data as part of a broader detection and validation ecosystem, rather than as a definitive representation of real-world risk.

7 Next steps

The FCA data sprint

The first practical use case for the data is as part of the upcoming data sprint. We want firms to use this data over the course of the sprint and come back together to share their findings.

We hope that through the use of the data, sharing findings and follow up discussions we can foster beneficial innovation in the detection of money laundering.

After the data sprint

When the data sprint has ended, the project will use participants' feedback from to improve the quality of the datasets, so they continue to reflect the complexity of financial crime.

We're also considering broadening access to the datasets. For now, our sandbox is the right environment but in the longer term, scaling access across industry will require careful governance, privacy controls, and alignment with international standards.

If this happens, the supporting technical documentation, governance arrangements, and disclosure of limitations should evolve alongside it. This will help maintain trust in how the dataset is interpreted, used, and assessed over time.

Further challenges for future consideration could include:

- Typology expansion: Which additional typologies and behavioural variations should we prioritise for the next iteration of the datasets?
- Access and governance: How should we manage access once the datasets move beyond the sandbox? Options could include licensing frameworks, usage agreements or integration into regulatory testing regimes.
- Evaluation standards: What metrics and validation protocols should we use so that results from sandbox experiments can be compared, shared, and trusted?

8 Conclusions

By creating a resource that is both statistically faithful and privacy-preserving, we hope the project will enable firms to experiment with new and emerging approaches to detecting money laundering in a safe and controlled environment.

It will provide a credible and regulator-supported way to demonstrate the effectiveness of their approaches without privileged access to live banking data. This will level the playing field and foster competition, helping us in the fight against financial crime.

We hope that this project can set a precedent not only for AML innovation, but for how regulators and industry can collaborate to address some of the most persistent challenges in financial services.

If you wish to apply for the data sprint, you can do so here: https://events.fcainnovation.co.uk/Synthetic_AML_Solutions_Sprint),

Applications close on 26 April.

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk