

SYNTHETIC DATA

The New Way to Do Research

2026 guide for researchers and innovators



Paphichaya Amphawanwong
Demand Generation Specialist
Questionpro Thailand

Synthetic Data: The Data Approach Reshaping Research

What Researchers and Business Leaders Need to Know Before Using It

Data is the foundation of research and innovation today. But as data becomes more valuable, privacy and security concerns matter more too. Synthetic data has become one of the most talked-about topics in market research — as AI plays a bigger role in the research process, teams are looking for faster insights and lower costs during exploration.

This article covers the essentials on synthetic data, clearly and concisely.

1. What Is Synthetic Data?

Synthetic data is information generated by algorithms, not collected from real people or real events. It's designed to closely mirror the statistical patterns, relationships, and structure found in real data, without containing any actual personal or sensitive details.

One important distinction: synthetic data is not the same as taking real data and modifying or hiding it (data masking). It's built entirely from scratch, using techniques like machine learning, simulation, or rule-based algorithms.

Three key characteristics

1. Algorithmically generated — not derived from real individuals, so it can't be traced back to anyone
2. Statistically similar — preserves the same trends, distributions, and correlations as the original data, so it retains real analytical value
3. Privacy-protecting — since no real personal data is present, both legal and ethical risk are significantly reduced



Example: A hospital wants to share patient data with an external research team, but privacy law stands in the way. By generating synthetic patient records that reflect the same trends as the real data, researchers can analyze and test hypotheses without putting any real patient's privacy at risk.

2. Its Role in Research, and How It Compares

Synthetic data isn't meant to replace traditional research. Some researchers see it as a **"third dimension"** of research alongside qualitative and quantitative methods used to explore ideas before committing to a full study. It helps test whether a hypothesis is worth pursuing, simulate early expert interviews, and refine survey questions before fielding them to real respondents.

- **Worth remembering:** Synthetic data results should be treated as a **directional indicator, not final proof**.

It's also useful for simulating rare or sensitive scenarios that are hard to capture with real data. For example, a university team studying mental health trends used synthetic survey data to protect participant identities, letting them share the dataset with researchers worldwide without any privacy risk.

Compared to Real Data and Data Masking

Criteria	Real Data	Data Masking	Synthetic Data
Source	Real events/users	Modified real data	Fully generated from scratch
Privacy risk	High	Reduced, but gaps remain	Very Low
Flexibility	Limited	Limited by masking method	High

Bottom line: Work that demands full accuracy like regulatory reporting, still needs real data. But when privacy is the blocker, synthetic data is far more flexible.

3. How It's Generated

Three main approaches:

- **Rule-based:** Uses predefined formulas. Straightforward, but lacks real-world complexity.
- **Statistical Modeling:** Analyzes real data's averages and correlations, then generates new data matching those statistics.
- **Advanced Machine Learning:** Uses **GANs** (two neural networks competing until output is nearly indistinguishable from real data) and **VAEs** (encode-decode data to produce new samples that preserve complexity).

The 5-step process: Analyze real data → Train the model → Generate new data → Validate against real data → Deploy

The core challenge: Balancing realism with privacy. Data too close to the original risks overfitting and unintentional exposure of sensitive information. Complex data types — images, audio, text — are harder to synthesize and require significant computing resources.

4. Limitations and the Business Case

Limitations to know before using it

- **Can lose rare, subtle details** — sometimes the most valuable insights in research
- **Risks inheriting bias** from the source data, unless actively checked during generation
- **Not suited for work requiring full accuracy**, such as legal evidence or final measurement

Recommendation: Use it as a supplement, always validated against real data, not a full replacement.

Four business reasons to care

1. **Enables cross-organization and cross-border data sharing**, without hitting privacy law barriers
2. **Provides access to rare samples or events**, without waiting for them to occur or risking ethical issues
3. **Cuts cost and time** for testing early hypotheses, before investing in full real-data collection
4. **Reduces legal risk** as GDPR, PDPA, and similar laws grow stricter every year

Summary

Synthetic data is algorithmically generated information designed to mirror the statistical properties of real data, without retaining any real personal information. It lets organizations share data across borders, study rare scenarios, cut costs, and reduce legal risk. But it has clear limits, bias, lost detail, and unsuitability for work demanding full precision. The best approach: use it as a complement to qualitative and quantitative research, not a replacement.