

aindo

# Synthetic data

The new global standard  
for data access and sharing

Daniele Panfilo, PhD  
Co-founder & CEO @ Aindo

1.

# Synthetic Data: a new paradigm in data access



# The problem

60% to 73% of all potential data value is currently not realized

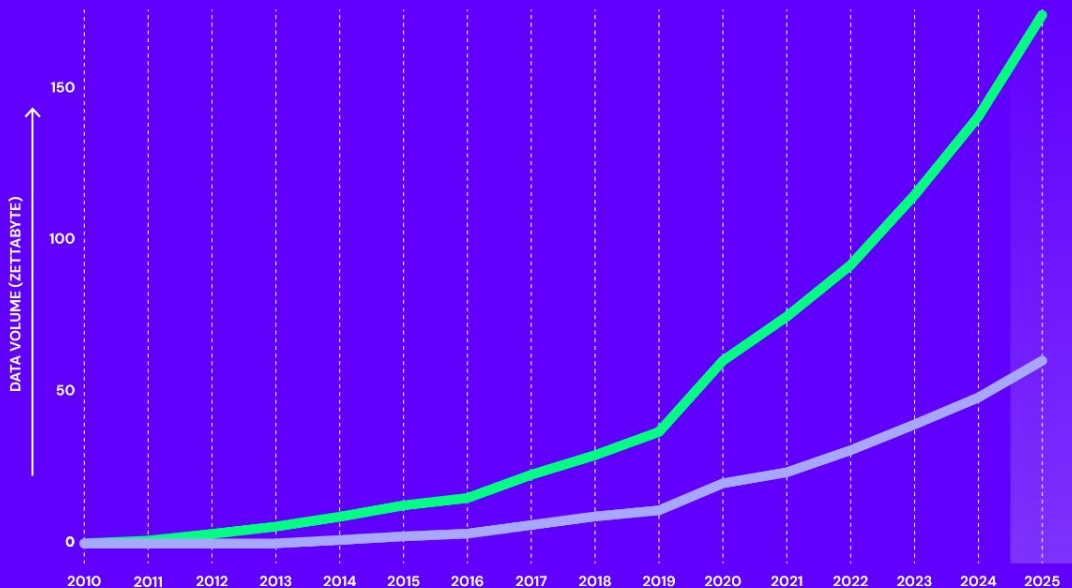
The data value gap



valorized data



available data



(Source: Statista; Accenture)

## Core obstacles

- **Data distribution**

Data is stored in different formats and systems, making it difficult to access and integrate.

- **Formatting**

Data is often collected in unstructured, unannotated formats.

- **Privacy**

Privacy legislation (GDPR, CCPA, DP) hinder data exchange and value extraction.

- **Incompleteness**

Data may contain crucial gaps, leading to inaccurate analytics and models

- **Bias and unfairness**

Datasets often underrepresent specific fragments of society, resulting in biased AI models



# What are synthetic data

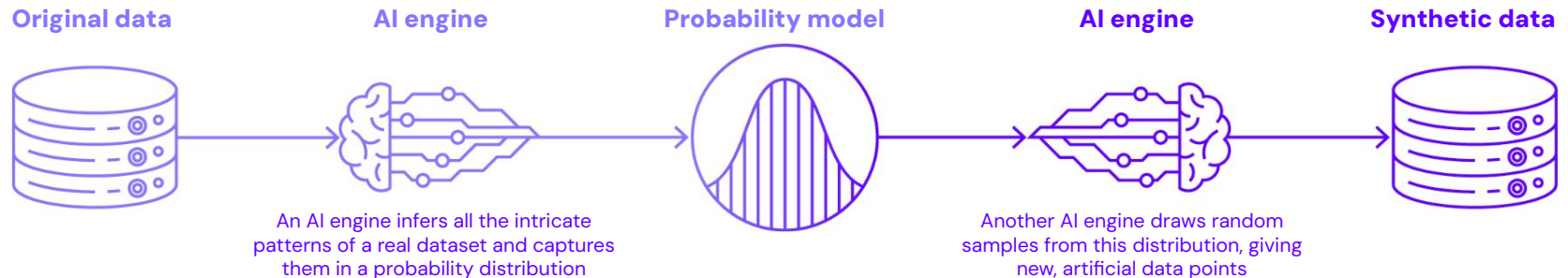
1

**Synthetic data** are **constructed algorithmically**, they are not collected empirically

2

Through advances in AI, synthetic data are **indistinguishable from real data**, yet **void of personal information**

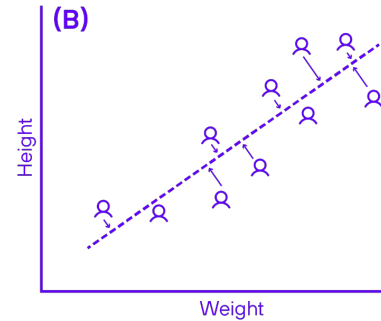
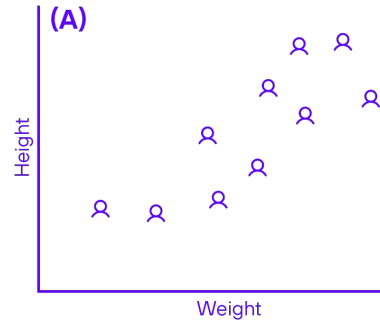
## Under the hood



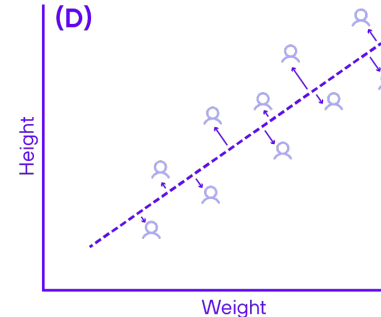
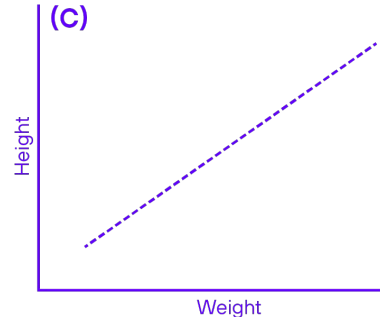
# Synthetic Data Replicate Real Patterns

Name	Height (cm)	Weight (kg)
John	181	77
⋮	⋮	⋮
Elsa	164	59

Real data



● Real  
● Synthetic  
--- Model



Name	Height (cm)	Weight (kg)
Tim	176	72
⋮	⋮	⋮
Luise	171	63

Synthetic data

# Synthetic data: quality assurance



## Privacy

How well does synthetic data use protect the privacy of real data subjects?



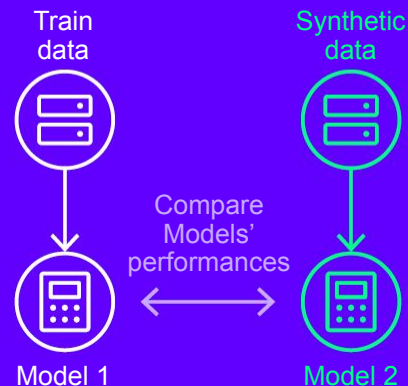
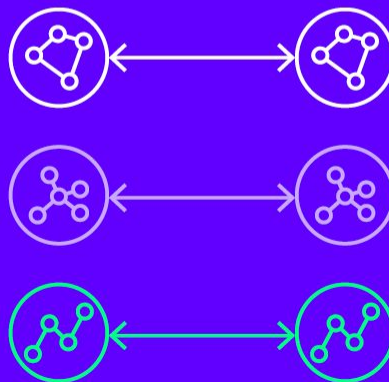
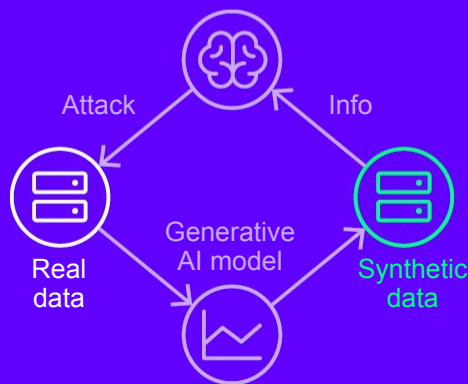
## Fidelity

How well does synthetic data preserve real statistical patterns?



## Utility

How useful are synthetic data in AI training?



1.

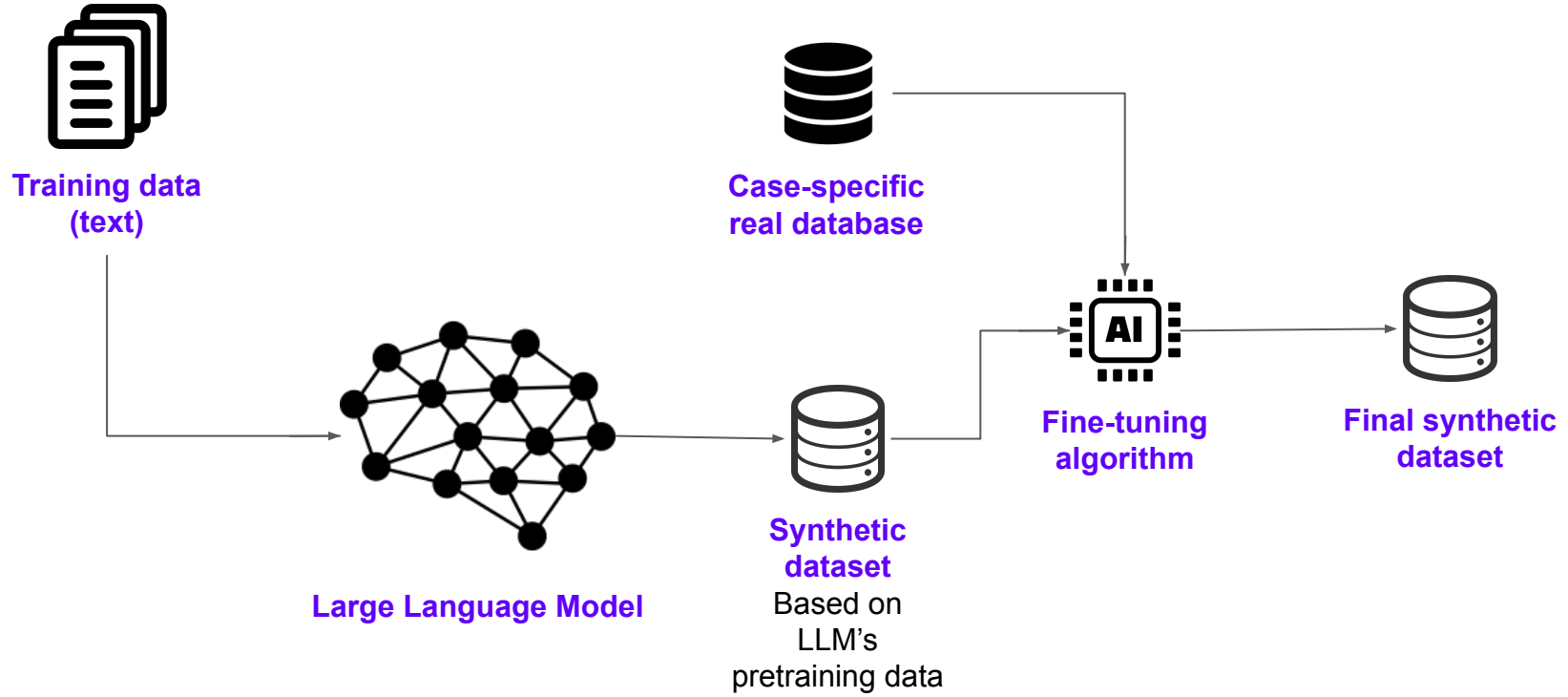
# Synthetic Data: Technological Evolution and State-of-the-Art

# Synthetic Data: Technological Evolution

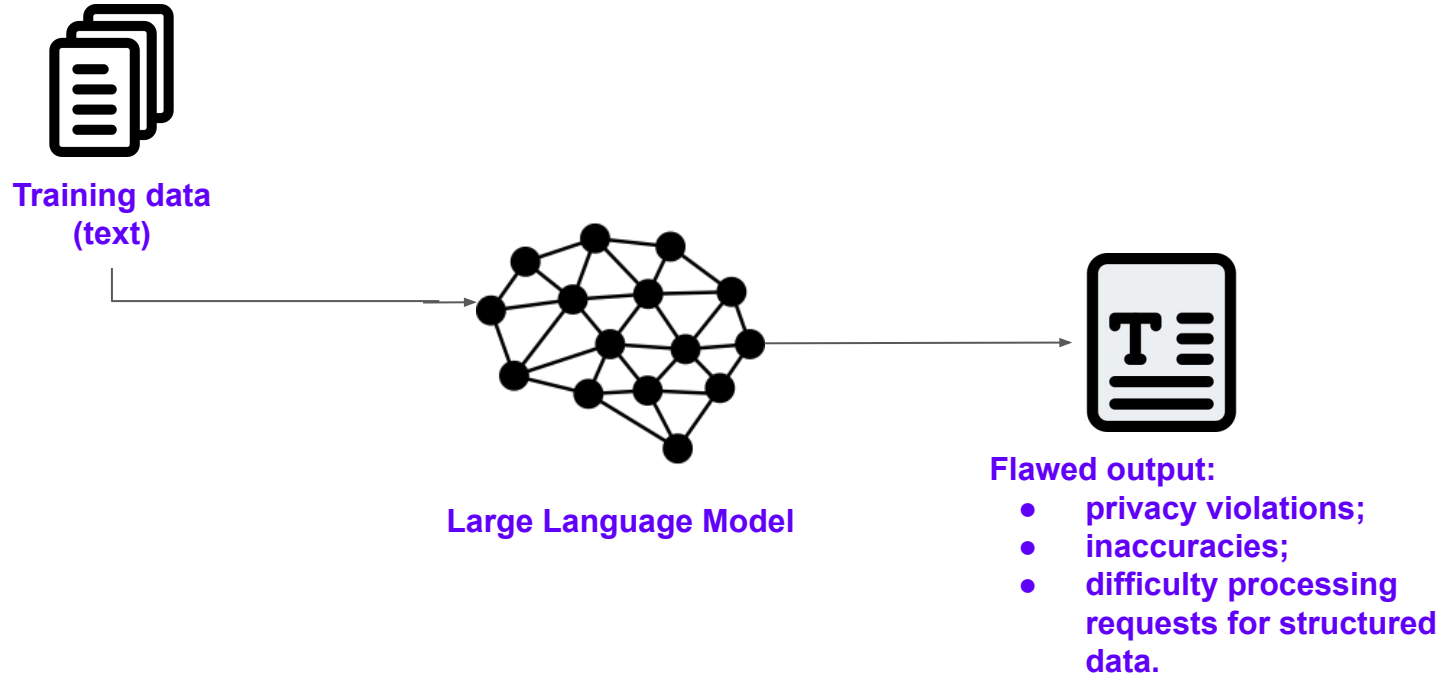
Year	Method	Technological advance(s)	Ideal data type(s)	Caveat(s)	Influential publication(s)
<b>Pre-2014</b>	Statistical, rule-based, and simulation tech.	<ul style="list-style-type: none"> <li>Initial synthetic data technologies;</li> </ul>	Tabular data	Typically require domain expertise, lack of automation	
<b>2014</b>	Generative Adversarial Networks (GANs)	<ul style="list-style-type: none"> <li>Initial gen AI-based generation tech;</li> <li>Automated, general-purpose models;</li> </ul>	Image data, structured data after preprocessing	Limited quality for structured data; subject to mode collapse	<a href="#">Goodfellow et al. (2014). Generative Adversarial Networks.</a>
<b>2014</b>	Variational Autoencoders (VAEs)	<ul style="list-style-type: none"> <li>Initial gen AI-based generation tech;</li> <li>Automated, general-purpose models;</li> <li>Strong performance (privacy/utility).</li> </ul>	Image data, structured data after pre-processing	Decent quality, but fails to meet domain-specific expectations	<a href="#">Kingma and Welling. (2014). Auto-Encoding Variational Bayes.</a>
<b>2017</b>	Transformer models	<ul style="list-style-type: none"> <li>Very strong performance due to conditional value assignment.</li> </ul>	Text (tech behind LLMs); images; structured data	Relatively computationally expensive	<a href="#">Vaswani et al. (2017). Attention is All You Need.</a>
<b>2023</b>	LLMs/foundation models	<ul style="list-style-type: none"> <li>Benefit of pretraining: reduced computational burden.</li> </ul>	Text; increasingly applied for structured data.	Fine-tuning for structured data complex	<a href="#">Sui et al.. (2024). Table Meets LLM.</a>



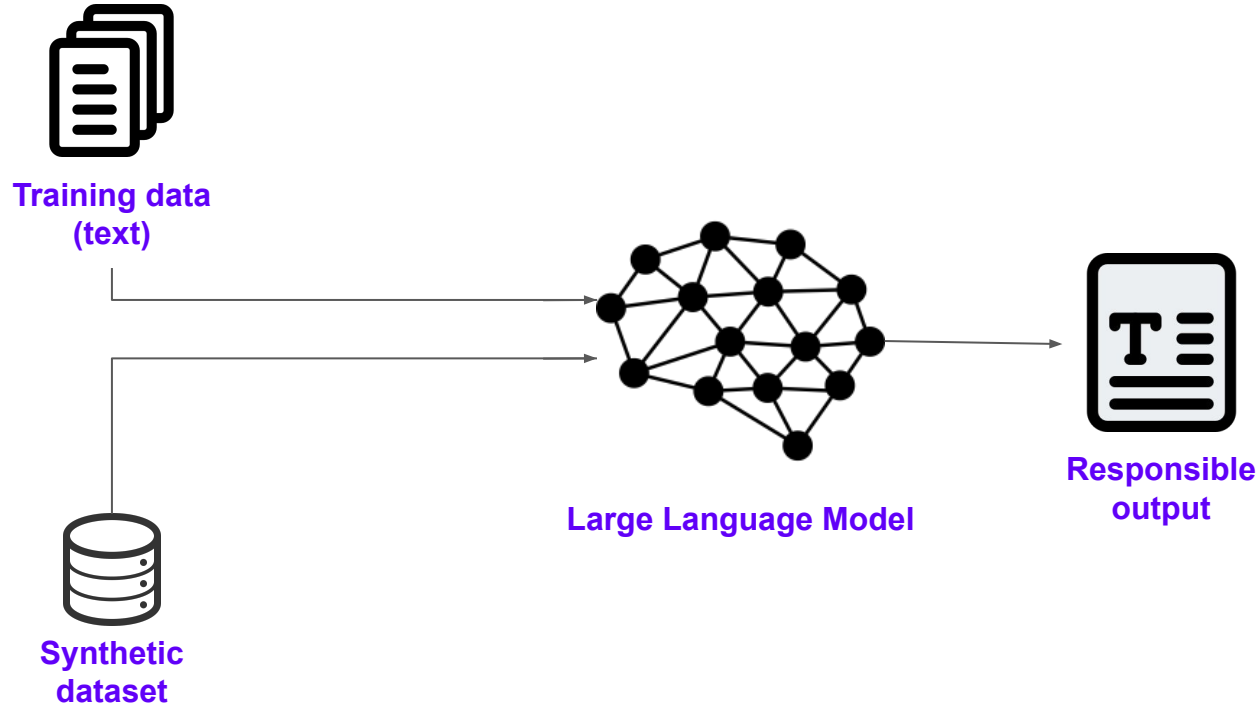
# Synthetic Data and LLMs: Symbiosis



# Synthetic Data and LLMs: Symbiosis



# Synthetic Data and LLMs: Symbiosis

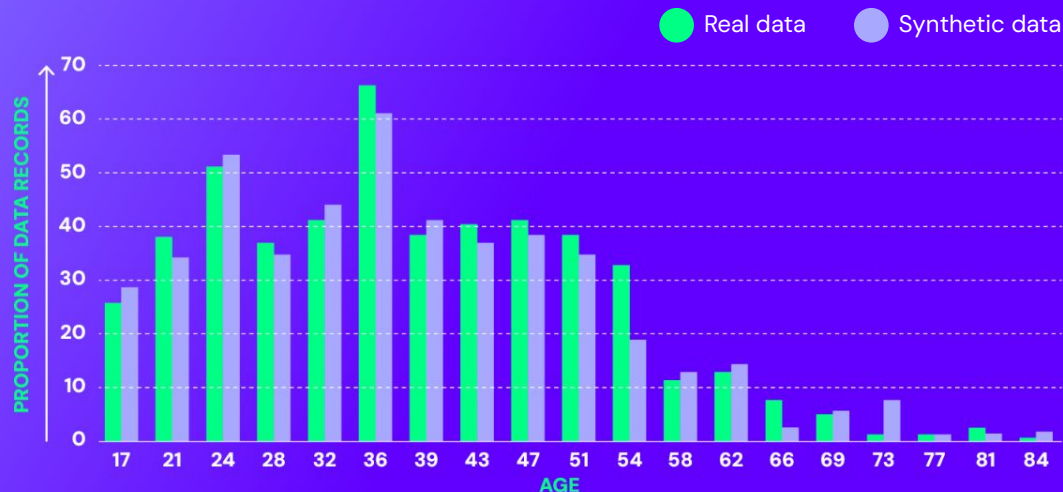


## 2.

# Synthetic Data: Fidelity and Utility



# Synthetic Data preserves realism needed for AI and analytics (Fidelity)



Comparing **marginal distributions**:  
how well do distributions in synthetic  
data resemble those of the real data?

Real data



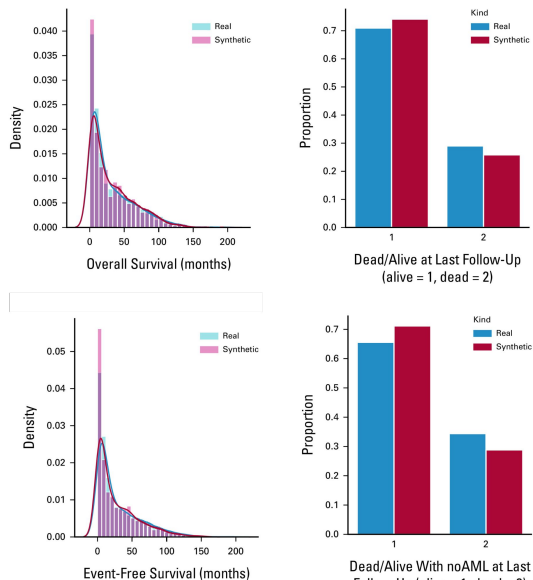
Synthetic data



Comparing **correlation matrices**: how well are  
the relations between  
variables preserved in  
the synthetic data?

# Synthetic Data Quality in recent studies

## Fidelity



Real (blue) and synthetic (red) survival features in a recent confrontational study on myelodysplastic syndromes (MDS) and AML.

Source: D'Amico, S. et al., (2023) "Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology"

## Utility

TABLE 2. Utility Comparison Using Three Generative Models

Data Set	Sample Size	SEQ			GAN			VAE		
		Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap
REaCT-HER2+	48	1	1	0.77	1	1	0.88	1	1	0.94
REaCT-G/G2	401	1	1	0.91	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	1	1	0.67
REaCT-ILIAD	218	1	1	0.99	1	1	0.85	1	0	0.74
REaCT-ZOL	211	1	<sup>b</sup>	0.98	1	<sup>b</sup>	0.88	0	<sup>b</sup>	0.61
REaCT-BTA	230	1	1	0.85	1	0	0.68	1	0	0.72
CCTG MA27	7,576	1	1	0.90	1	1	0.62	1	1	0.82
SWOG 0307	6,097	1	1	0.93	1	0	0.50	1	1	0.95
NSABP B34	3,323	1	1	0.93	1	1	0.83	1	1	0.61

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

<sup>a</sup>Training the generative model failed.

<sup>b</sup>The analysis is descriptive and hence decision agreement does not apply.





Utility assessment of **three distinct generative models** and **eight distinct medical datasets**. Estimates, decisions, and confidence intervals aligned well in most cases, even for very small sample sizes.

Source: El Kababji, S. et al., (2023) "Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets"

# Monzino: Synthetic data contrastive study (Utility)



<b>Objective</b>	<input type="checkbox"/> Compare real and synthetic cohort to predict changes in <b>LDL-C levels</b> and achievement of <b>LDL-C goals</b> ;
<b>Data</b>	<input type="checkbox"/> Sample size: <b>n≈700</b> (per cohort real/synthetic). <b>Variables</b> include: gender, age ranges, intensity of care, ...
<b>Methods</b>	<input type="checkbox"/> Aindo <b>SDK integration</b> in data pipeline; <input type="checkbox"/> <b>Evaluate synthetic data fidelity:</b> <ul style="list-style-type: none"> <li>• <u>Jensen-Shannon divergence</u> between synthetic and real categorical attributes;</li> <li>• <u>Kolmogorov-Smirnov (KS) tests</u> to compare distributions of synthetic and real numerical attributes.</li> </ul> <input type="checkbox"/> <b>Evaluate synthetic data utility:</b> Generalized Linear Model built on synthetic and real data: <ul style="list-style-type: none"> <li>• <u>Comparison of odds ratios</u> at 95% confidence interval;</li> <li>• <u>Stepwise variable selection</u>: are the same variables selected, with same correlations, in synthetic and real data at each step? Evaluated with cross-validation.</li> </ul> <input type="checkbox"/> <b>Evaluate synthetic data privacy metrics</b> through Aindo's SOTA built-in metrics.

<b>Results</b>		<b>Fidelity assessment</b>	<input checked="" type="checkbox"/> Significantly small Jensen-Shannon divergence obtained; <input checked="" type="checkbox"/> KS tests at 95% confidence fail in <5% of cases;
		<b>Utility assessment</b>	<input checked="" type="checkbox"/> Comparable odds ratios between in synthetic and real data; <input checked="" type="checkbox"/> Same variables were selected with comparable correlations in stepwise variable selection;
		<b>Privacy assessment</b>	Score of >95%, indicating absence of severe risks
		<b>Time savings estimate</b>	>66%, from ~9 months to ~3 months

## 3.

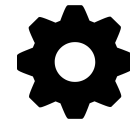
## Synthetic Data: Privacy



# Importance of Privacy Quantification Underestimated



Source: Kaabachi, B. et al., (2023). [Can We Trust Synthetic Data in Medicine? A Scoping Review of Privacy and Utility Metrics.](#)

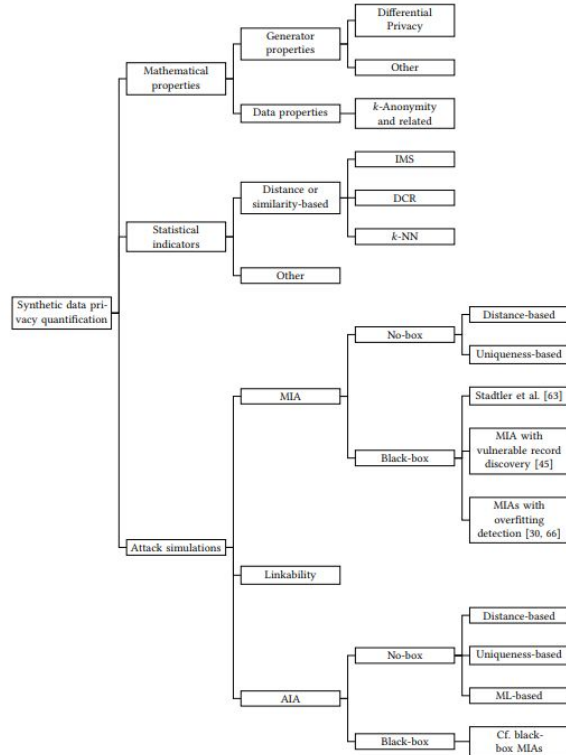


# Technical Privacy Measures

<p><b>Control privacy</b> Deploy privacy mechanisms such as differential privacy</p>	<ul style="list-style-type: none"> <li>• <b>User-controlled</b>, context-dependent threat tolerance;</li> <li>• Provable guarantees, <b>independent of attacker's technology</b> (GDPR recital 26).</li> </ul>
<p><b>Measure privacy</b> Use statistical privacy metrics to quantify potential risks</p>	<ul style="list-style-type: none"> <li>• <b>Gauge weaknesses</b> that can be exploited in attacks;</li> <li>• Guarantees <b>independent of attacker's technology</b> (GDPR recital 26);</li> <li>• <b>No assumptions</b> about adversaries' technologies or deployment context.</li> </ul>
<p><b>Test privacy</b> Conduct deliberate attacks to ensure these do not succeed</p>	<ul style="list-style-type: none"> <li>• Explicitly model specific implementations of the <b>WP29 attacks</b>;</li> <li>• Rely on specific <b>threat models</b>: what tools does the adversary have at their disposal? (GDPR Recital 26: tools "reasonably likely" to be used).</li> </ul>

Further reading: Boudewijn et al., (2023). [Privacy Measurement in Tabular Synthetic Data: State of the Art and Future Research Directions](#).

# Need for Standardized, Auditable Privacy Criteria



A plethora of privacy quantification methods exists:

- Different frameworks (indicators; provable guarantees; empirical tests);
- Different risk factors (similarity; uniqueness; adversarial ML);
- Different threat models: what can adversaries use?

There's a strong need for **standardization**, in light of current and upcoming privacy legislation.

# Organizational Privacy Measures



Local versus cloud deployment	<ul style="list-style-type: none"><li>• <b>Local deployment</b> of synthetic data technology avoids exposure of sensitive data outside of its original environment;</li><li>• Secure (private) <b>cloud deployment</b> is possible with additional cybersecurity steps.</li></ul>
Limit generator access	<ul style="list-style-type: none"><li>• Adversaries can use any form of <b>additional information</b> in conducting attacks. This includes <b>access to the generative model</b>.</li><li>• <b>Closed-source generators</b> are therefore preferable from a privacy viewpoint.</li></ul>
Certification	<ul style="list-style-type: none"><li>• <b>Europrivacy™/® certification</b> guarantees that generator processing meets GDPR requirements;</li><li>• Synthetic data providers should further adhere to <b>recognized standards</b> such as ISO/IEC 27001;</li><li>• ISO/IEC 42001 on AI Management Systems aligns closely with the AI Act and ensures <b>responsible use of AI systems</b>.</li></ul>
Deploy the technical measures	<ul style="list-style-type: none"><li>• Identify and account for potential risk factors.</li></ul>





# Aindo: Europrivacy™/® certified



**Building Trust and Confidence  
in Privacy and Data Protection**

[www.europrivacy.com](http://www.europrivacy.com)

Comprehensive validation of data processing activities' **compliance with the GDPR** and national legislation.

Aindo is the **first and only Europrivacy™/® certified synthetic data provider**.

## What this means



Reduced Legal and Financial Risk



Privacy practices recognized across EU and EEA

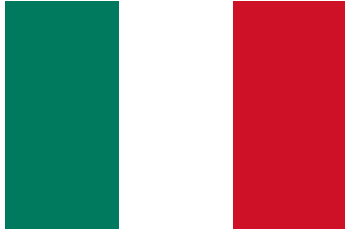


Compatibility with other standards



Future-proof data handling and ongoing commitment

# Beyond the GDPR: synthetic data in Italy



Proposed legislation  
on AI - DDL n. 1146  
positions Italy as a  
leader in AI-driven  
healthcare research.

## Article 8: AI-driven research = public interest

---

- §1 The processing of personal data **for AI systems** in healthcare research is a "relevant public interest"  
→ **This eliminates the need for case-by-case justification that previously created barriers to innovation.**
- §3 Explicitly authorizes "processing for the purpose of anonymization, pseudonymization or synthesis of personal data", including sensitive data (special categories ex Art. 9 GDPR)  
→ **No prior approval required for anonymization activities, including synthetic data generation.**
- §4 AGENAS, under the guidance of the Italian Data Protection Authority (Garante), is set to issue and maintain technical standards for synthetic data and anonymization processes  
→ **This ensures harmonized and future-proof governance.**

**Davide Ruffo, LL.M**  
**Chief of institutional relations**  
**+39 3391471728**  
**[davide.ruffo@aindo.com](mailto:davide.ruffo@aindo.com)**

[illegible]