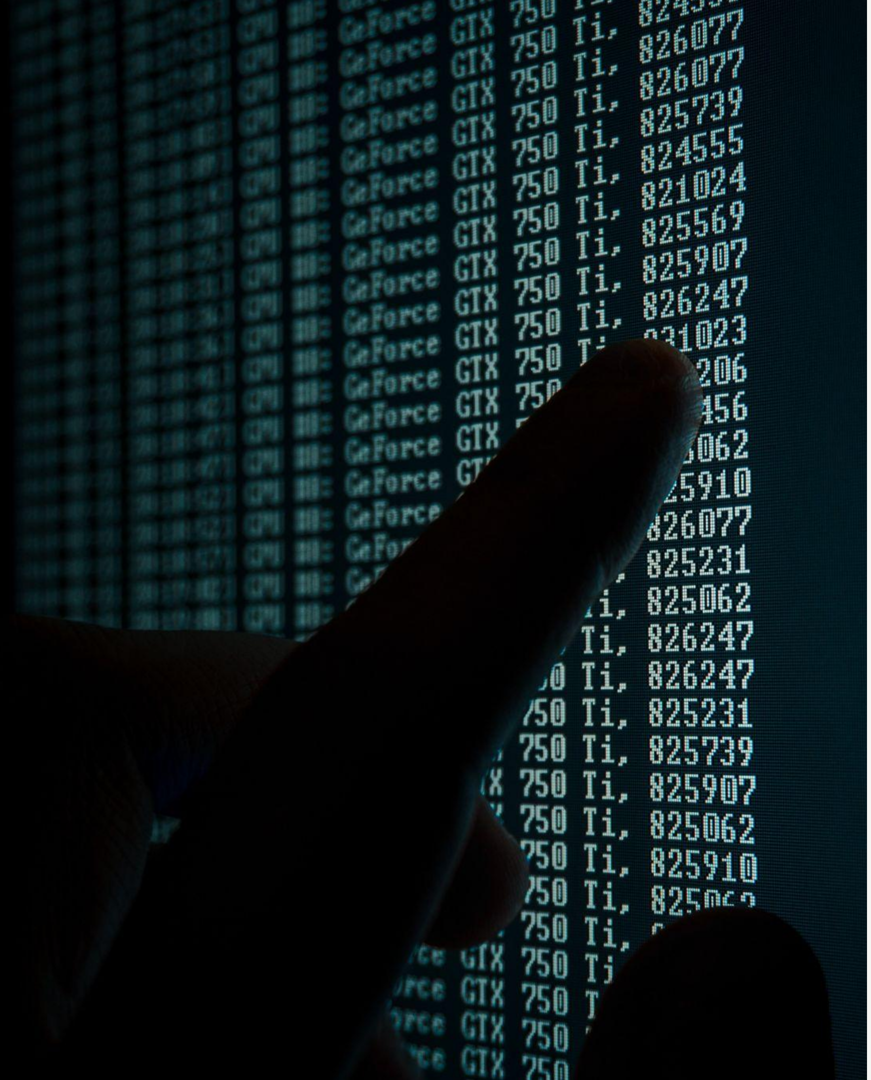


Synthetic data as a novel anonymization technique: generation and evaluation approaches

Marta Batlle López, Data Scientist (Roche)

16 May 2024

The core problems of sharing healthcare data

It is **challenging to access and share valuable healthcare data**

But high-quality data is still needed to **drive innovation**

In response, many **anonymisation techniques and privacy-enhancing technologies** were born

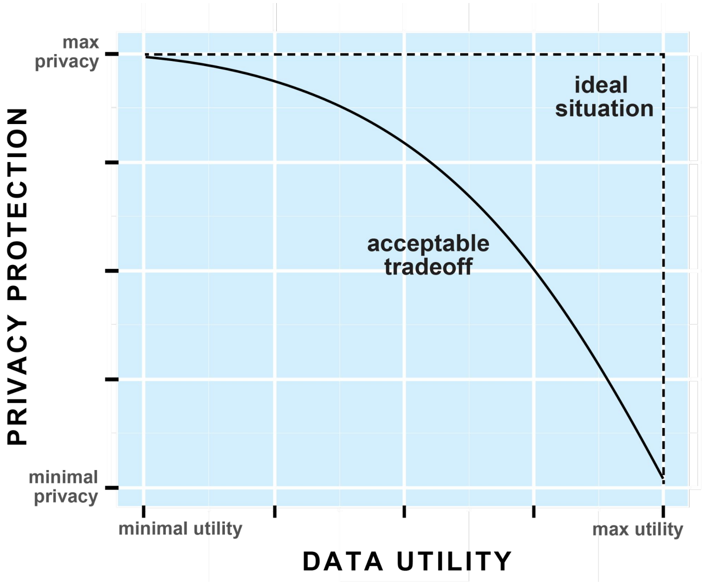
How to choose between privacy enhancing technologies?

The quality and privacy trade-off

Many anonymisation techniques and privacy-enhancing technologies **were born** to tackle the problem of data sharing

Some can **compromise the quality of the data or pose technical challenges**

We look for that technology or technique that allows us to find an **acceptable trade-off** between the quality of the data and the level of privacy protection



How to choose between privacy enhancing technologies?

Technology	Pros	Cons
Homomorphic Encryption	High security & confidentiality (data usable even if encrypted)	Computationally intensive
Differential Privacy	Strong privacy guarantees (adds noise to data in a controlled way)	Data utility (may reduce accuracy of analysis)
Data Anonymization	Low risk data sharing (removes all PII data)	Data utility might be highly reduced
Pseudonymization	Easy data sharing (removes direct identifiers)	Risk of re-identification
Federated Learning	Collaborative learning (trains models on distributed data)	Complex coordination, potentially lower accuracy (aggregates)
Secure Multi-Party Computation (SMPC)	Strong Privacy for Joint Computations (Parties compute without revealing data)	Computationally intensive
Synthetic data	Private data sharing (generate artificial data based on real distributions)	Potentially lower utility of the data


StyleGAN2 (Karras et al.)

What is synthetic data?

Real dataset

	Age	Weight	...	Height
	20	60	...	170
	22	80	...	190
	...	...	...	...
	24	70	...	180
AVG	**22**	**70**	**...**	**180**
STD	**2**	**10**	**...**	**10**

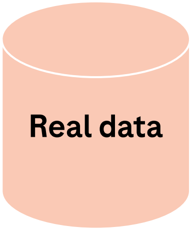
SDG algorithm

Synthetic dataset

Age	Weight	...	Height
19	85	...	195
22	70	...	180
...	...	...	...
25	55	...	165
22	**70**	**...**	**180**
3	**15**	**...**	**15**

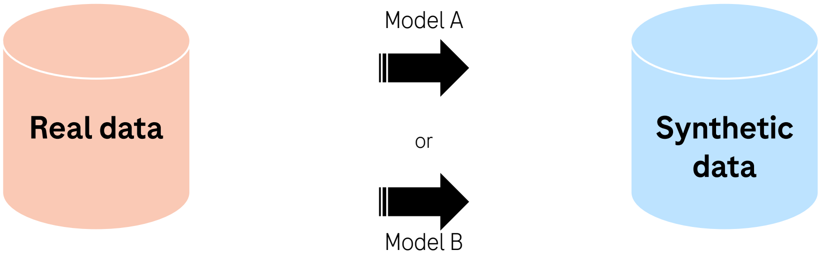
Similar mathematical significance

Synthetic Data Generation process

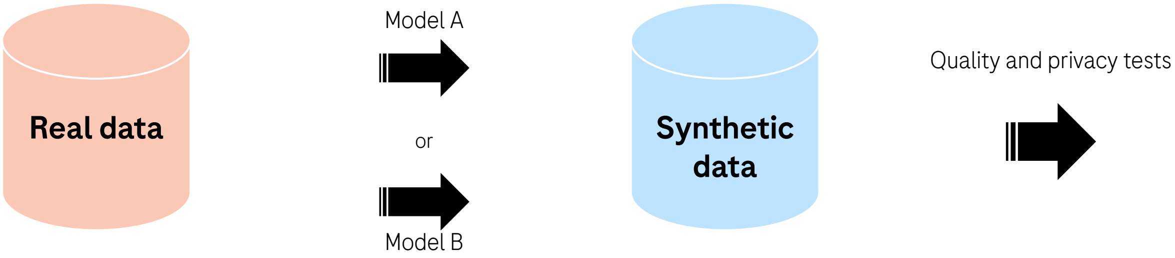


Synthetic Data Generation process

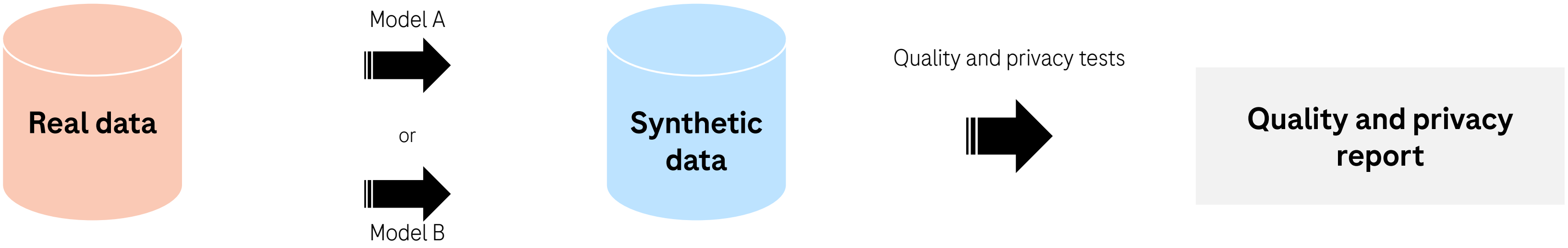
Synthetic Data Generation process



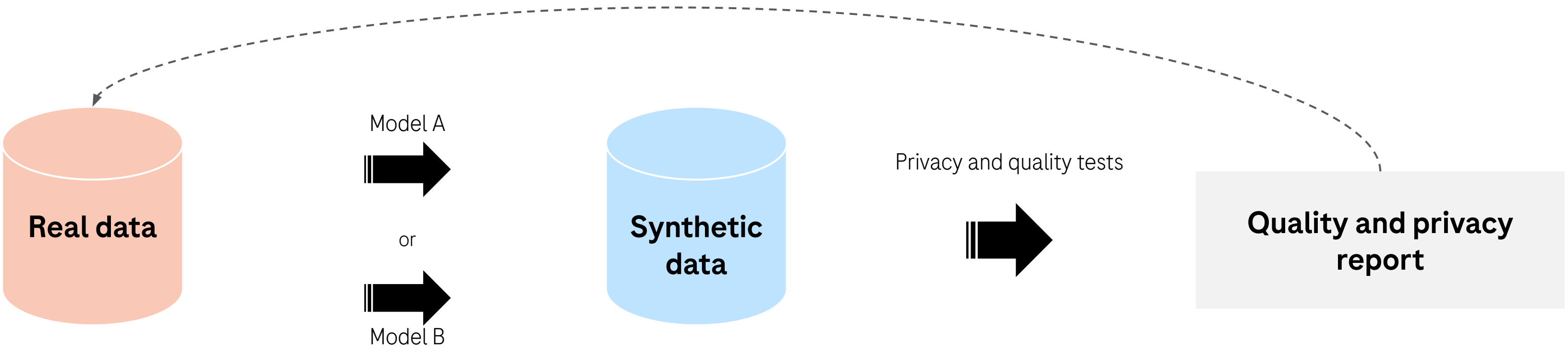
Synthetic Data Generation process



Synthetic Data Generation process



Synthetic Data Generation process

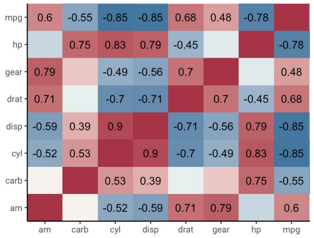


Generic utility evaluation metrics

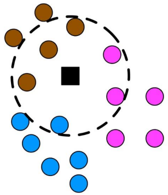
How do we assess the quality of the synthetic data?

Variable	N	Mean	SD
X	...	...	...
Y	...	...	
Z	...	...	...

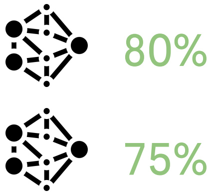
Feature distributions



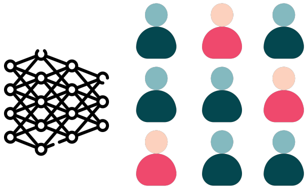
Feature correlations



Distance metrics



ML performance



Distinguishability

Measuring privacy risks

How do we assess the privacy of the synthetic data?

An effective anonymisation prevents an attacker from:

1. **Singling out an individual** in a dataset
2. **Linking two records** within a dataset (or between two separate datasets)
3. **Inferring any information about individuals** in such dataset

Original sample

Age	Gender	BMI	Hypertension
58	F	32.4	True
66	M	25.4	False
35	M	27.8	True

Synthesis

Synthetic sample

Age	Gender	BMI	Hypertension
35	M	28	True
29	F	34.4	True
65	M	30.8	False

Measuring privacy risks

How do we assess the privacy of the synthetic data?

Identity Disclosure involves the risk of inferring the true identity of an individual from synthetic data

Adversary gains information on the individual through the **matched record**.

Membership Disclosure refers to the risk that an individual's presence in a dataset can be disclosed through the synthetic data.

Adversary gains information on the individual through the **membership in the dataset.**

Original sample

Age	Gender	BMI	Hypertension
58	F	32.4	True
66	M	25.4	False
35	**M**	27.8	True

Synthetic sample

Age	Gender	BMI	Hypertension
35	**M**	28	True
29	F	34.4	True
65	M	30.8	False

Applications of synthetic data

Higher quality of the synthetic data is required

"I want to solve imbalances in my data"

"I need to use data for secondary purposes"

"I want to augment the size of my data"

"I want to facilitate a data sharing process"

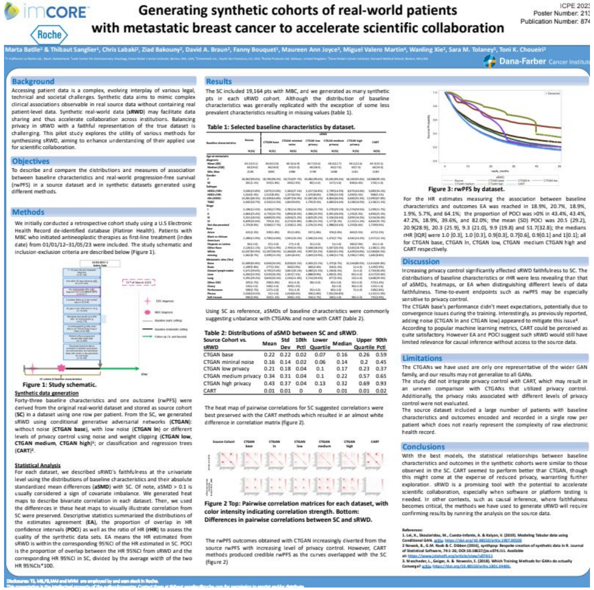
"I want to test our software, model, pipeline or platform"

"I want to start building my solution before getting data access"

"I want to test a software before contracting"

A real example of synthetic data sharing pilot

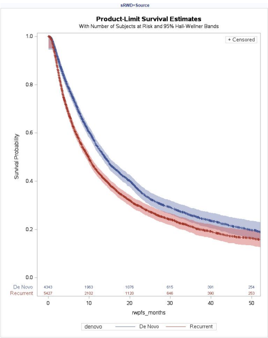
- Pilot on data sharing to accelerate collaboration with Dana Farber Institute (Harvard Medical School)
- Synthesis of sRWD from existing *analytical dataset* using different methods CTGAN or CART
    - ~10k MBC patients
    - Over 100 variables

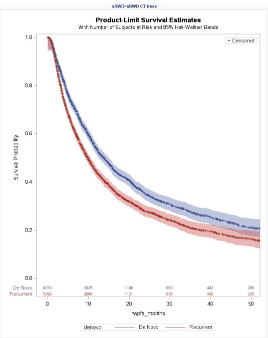


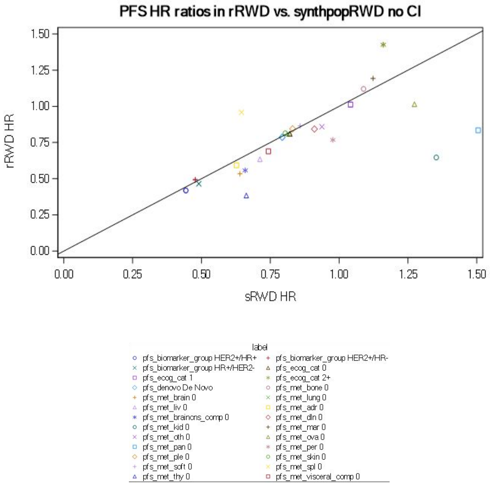
A real example of synthetic data sharing pilot

- A total of 9,770 pts with MBC were included in the SC and as many synthetic pts were generated in each sRWD cohort.
- Distributions of continuous and categorical variables were closely replicated
- Measures of association between baseline characteristics and outcomes were largely preserved
- CART outperformed CTGAN
- Dimensionality of dataset has a big impact on utility

			sRWD					
	Source		sRWD GANI		sRWD GANm		sRWD CART	
	N	(Percent)	N	(Percent)	N	(Percent)	N	(Percent)
biomarker_group								
HER2+/HR+	1,430	**(14.62)**	1,149	(11.74)	1,667	(17.04)	1,427	(14.59)
HER2+/HR-	669	**(6.84)**	904	(9.24)	711	(7.27)	661	(6.76)
HR+/HER2-	6,294	**(64.33)**	6,136	(62.71)	5,485	(56.06)	6,345	(64.85)
TNBC	1,391	**(14.22)**	1,595	(16.30)	1,921	(19.63)	1,351	(13.81)
bmi_cat								
1.Underweight	244	**(2.49)**	236	(2.41)	218	(2.23)	266	(2.72)
2.Normal	2,686	**(27.45)**	2,368	(24.20)	2,353	(24.05)	2,713	(27.73)
3.Overweight	2,766	**(28.27)**	2,718	(27.78)	3,767	(38.50)	2,702	(27.62)
4.Obese	3,339	**(34.13)**	3,778	(38.61)	2,595	(26.52)	3,335	(34.09)
5.Missing	749	**(7.66)**	684	(6.99)	851	(8.70)	768	(7.85)
groupstage								
0	1	**(<1.0)**	43	(<1.0)	60	(<1.0)	3	(<1.0)
I	1,006	**(10.28)**	1,104	(11.28)	1,065	(10.89)	1,032	(10.55)
II	2,219	**(22.68)**	1,820	(18.60)	2,309	(23.60)	2,230	(22.79)
III	1,518	**(15.52)**	2,036	(20.81)	1,908	(19.50)	1,508	(15.41)
IV	4,280	**(43.74)**	4,017	(41.06)	3,665	(37.46)	4,302	(43.97)
Not documented	760	**(7.77)**	764	(7.81)	777	(7.94)	709	(7.25)







Thank you

Doing now what patients need next