Come integrare nuove tecnologie nella pratica clinica

Alberto Zambelli

Oncologia, Papa Giovanni XXIII, Bergamo
Università degli Studi di Milano-Bicocca

Generative AI in Breast cancer

A use case

Generative AI in Breast Cancer

An evolving scenario

Explanatory and pragmatic trials drive BC clinical research and may impact with innovative practice-changing treatment opportunities

Evolving therapeutic landscape

⇩

need of **personalized treatments**

Role of RWD

RWD are complemetary to RCTs and critical to optimize treatment

EHRs are source of RWD (diagnosis, procedures, medications, lab results)

⇩

privacy and security

⇩

data scarcity

Artificial Intelligence

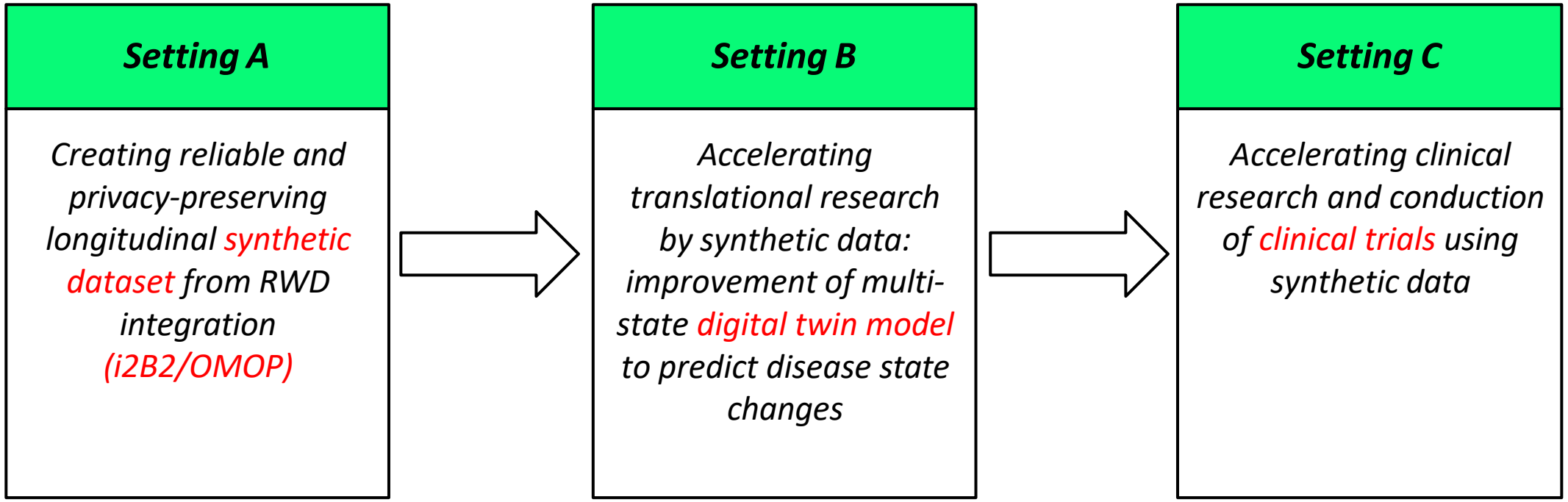
Need for large amount of data for training AI models

Synthetic data as a promising solution to overcome these barriers

⇩

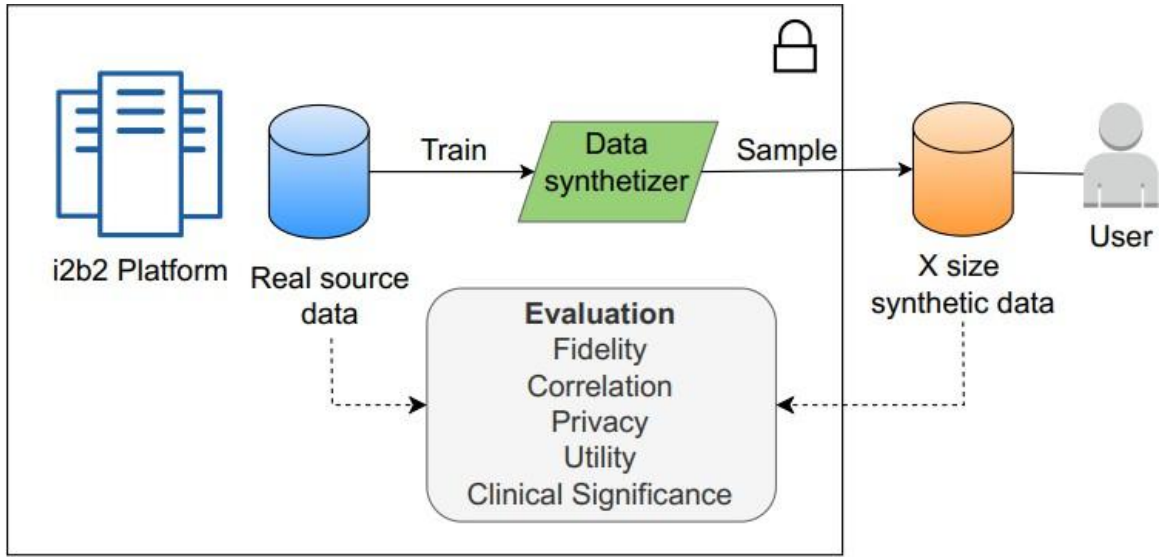
- preserve patient **privacy**
- facilitates secure **data sharing**
- **augment** existing datasets
- supports **ML model** development
- accelerates oncology **research**

Rationale


Setting A
Creating reliable and privacy-preserving longitudinal synthetic dataset from RWD integration (i2B2/OMOP)
Setting B
Accelerating translational research by synthetic data: improvement of multi-state digital twin model to predict disease state changes
Setting C
Accelerating clinical research and conduction of clinical trials using synthetic data

A) SD generation from RWD

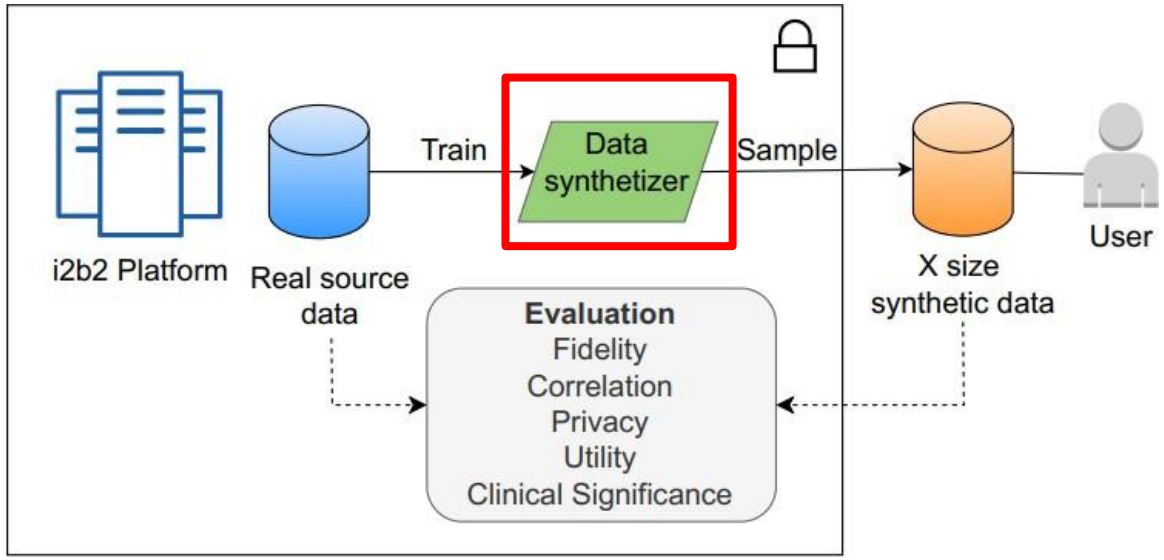
Setting A: *Creating reliable and privacy-preserving longitudinal synthetic dataset for data integration*



Objective	SD can replicate the complexity of RW longitudinal datasets while preserving patient privacy
Outcome	SD with generative AI models was suitable for integration with i2b2 ➔ augmented and anonymous dataset

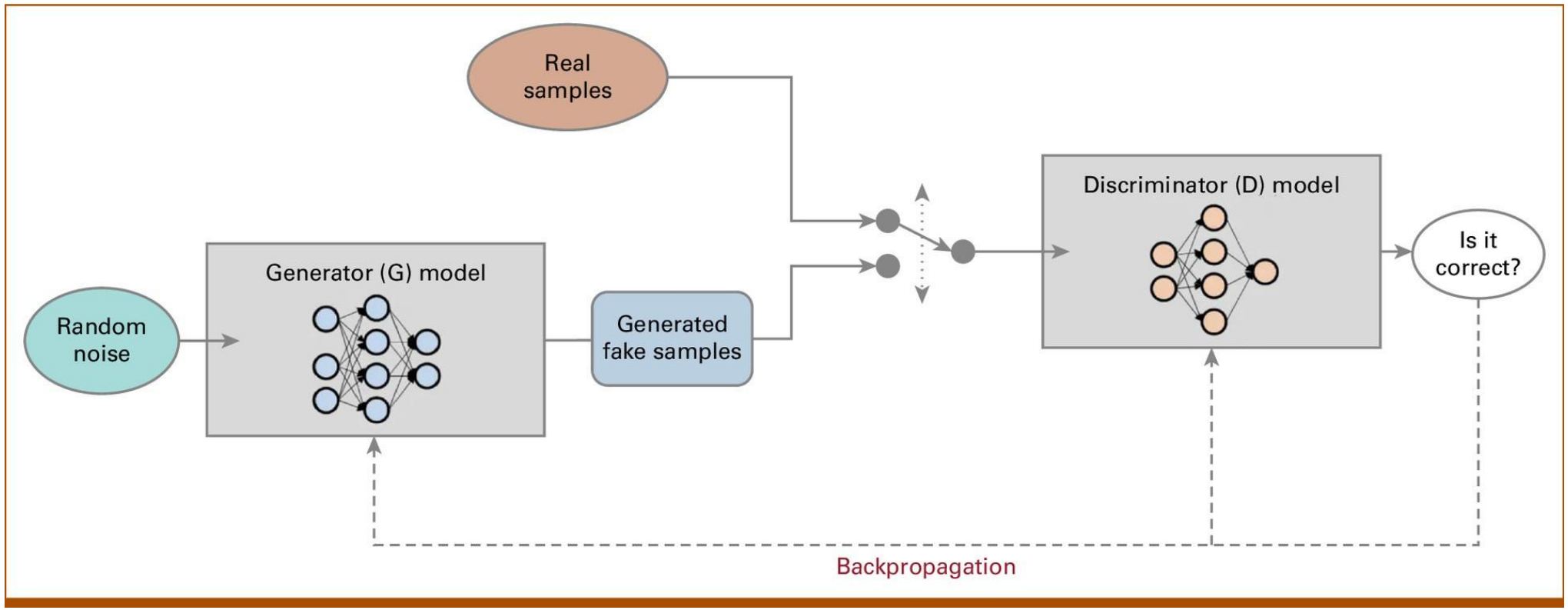
A) SD generation from RWD

Setting A: *Creating reliable and privacy-preserving longitudinal synthetic dataset for data integration*



Objective	SD can replicate the complexity of RW longitudinal datasets while preserving patient privacy
Outcome	SD with generative AI models was suitable for integration with i2b2 ➔ augmented and anonymous dataset

Data Synthetizer The Generative Adversarial Network (GAN)

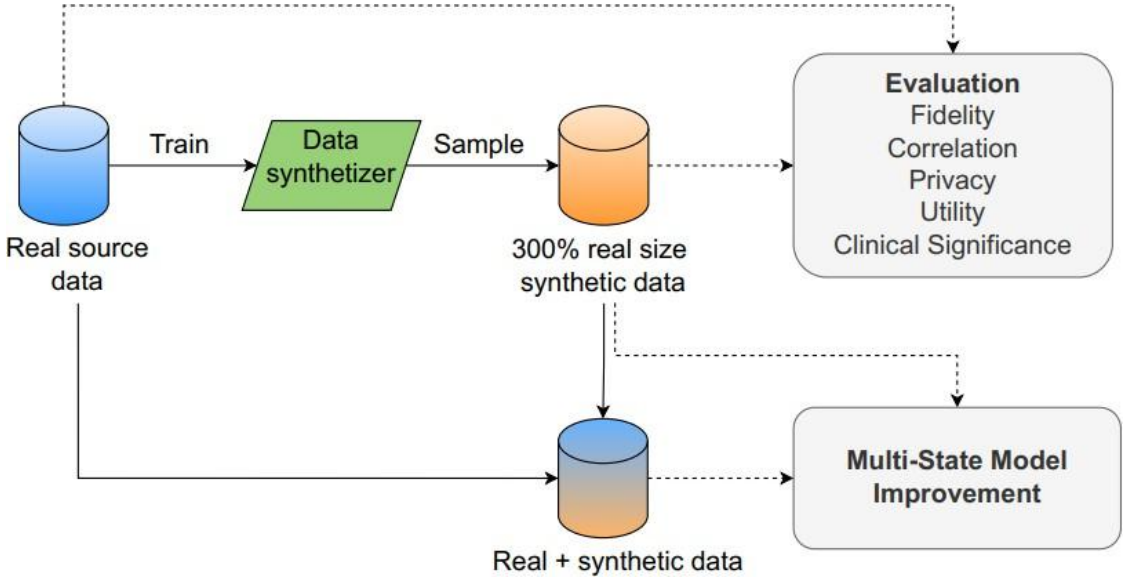


GAN functioning consists of two neural networks, the generator (G) and the discriminator (D).
The **generator** generates SD by taking random input and producing samples that closely resemble the real data.
The **discriminator** assesses whether the samples are real or fake and backpropagates the results to both networks to enhance their performance until the generator produces SD that is indistinguishable from real data.

F Jacobs et al JCO CCI 2024

B) Digital Twin model from SD

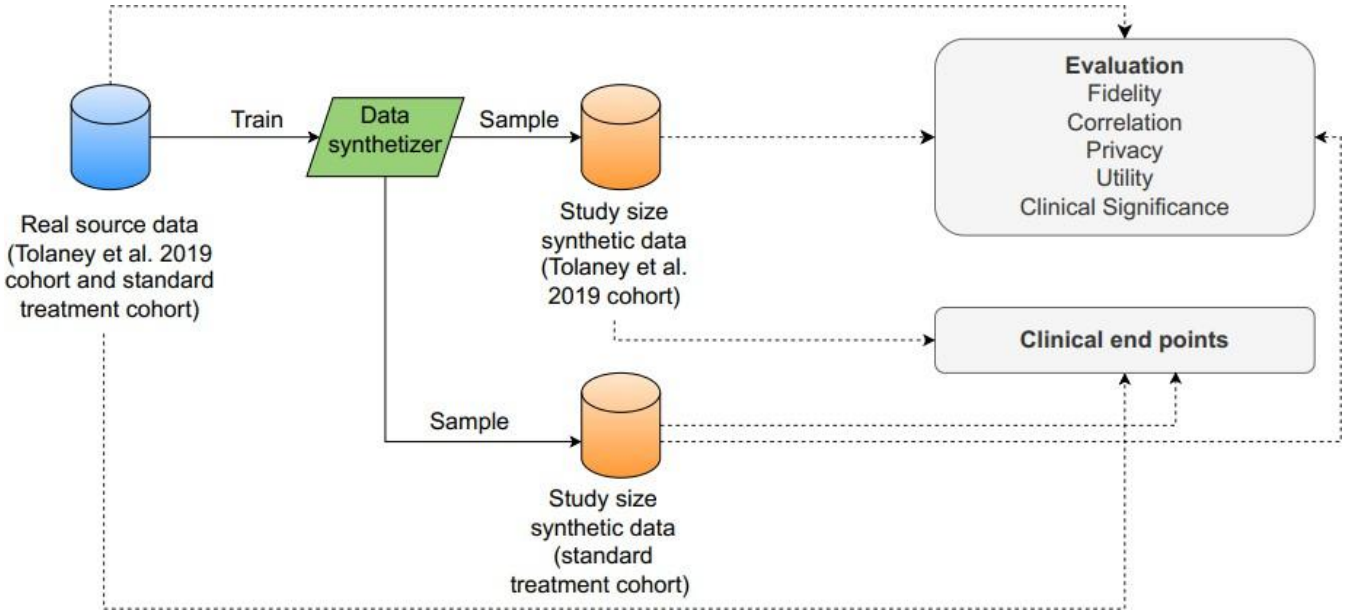
Setting B: *Accelerating translational research by SD: improvement of multi-state digital twin model to predict disease state changes*



Objective	Improve a multi-state digital twin model for predicting disease state transitions in BC
Outcome	Using SD for model augmentation enhanced performance metrics and supported better prediction of disease progression from states (early, recurrent, metastatic)

C) Clinical Trials with SD

Setting C: *Accelerating clinical research and conduction of clinical trials by using SD*



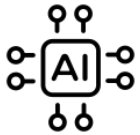
Objective	Evaluate SD utility in replicating clinical trial endpoints (*APT trial* for HER2+ N0 BC) and creating synthetic control arms
Outcome	SD replicated DFS outcomes, showing strong alignment with RW study results, and created a control arm that validated the study findings ➔ demonstrate feasibility for creating artificial control arms, optimizing trial cost and duration

Tolaney, NEJM 2015; Tolaney, JCO 2019; Tolaney, Lancet 2023

Methods

Study Population

i2b2 Platform and Cohort Identification



Generative Models for Longitudinal Synthetic Data

Synthetic Validation Framework

Experimental Set-up

Study Population

- 1052 BC pts
- Papa Giovanni XXIII Hospital

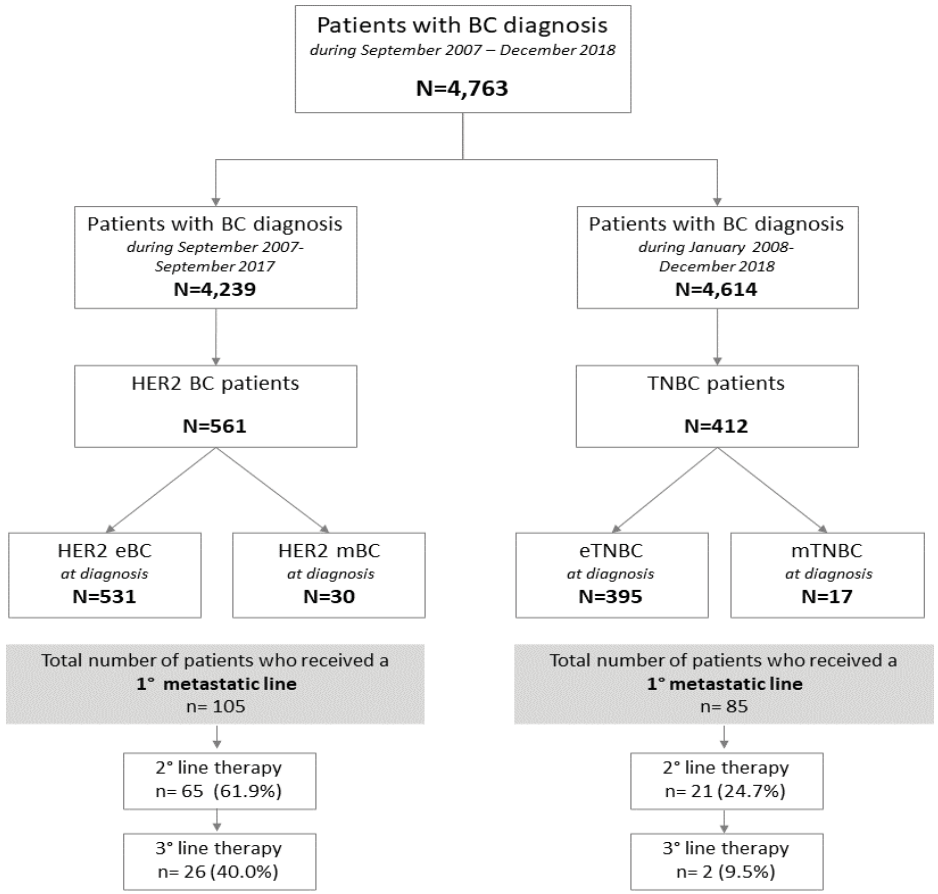
- HER2+ 58.8%
- TNBC 41.2%
- 10y period observation time
- Median follow-up 6.2y

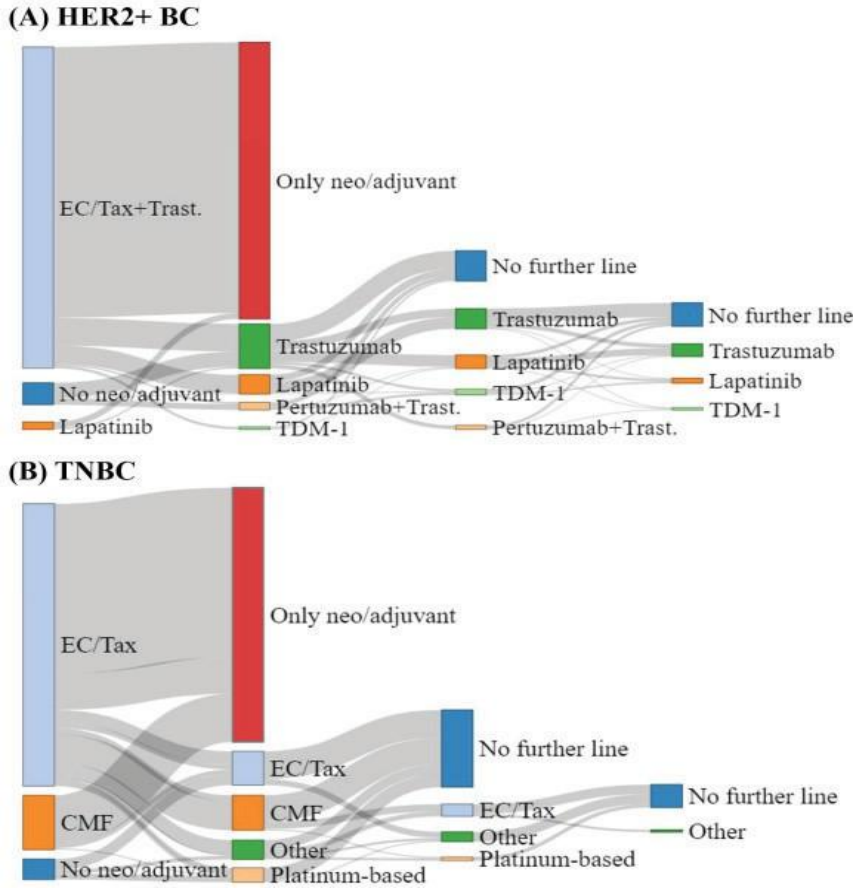
Information extracted

- *demographic data*
- *tumor characteristics*
- *treatments*
- *procedures*
- *survival*

Characteristic	All patients (N=1052)
	no. (%)
Age	
Median	50-59
<50	338 (32.1%)
50-59	257 (24.4%)
60-69	241 (22.9%)
>=70	214 (20.3%)
Grade	
I	30 (2.9%)
II	238 (22.6%)
III	511 (48.6%)
Disease stage	
Stage 1	169 (16.1%)
Stage 2	115 (10.9%)
Stage 3	71 (6.7%)
Stage 4	57 (5.4%)
HER2 status	
Positive	619 (58.8%)
Negative	433 (41.2%)
ER status	
Negative	410 (39.0%)
Low	49 (4.7%)
Positive	349 (33.2%)
PgR status	
Negative	605 (57.5%)
Low	132 (12.5%)
Positive	270 (25.7%)
T stage (TNM)	
T0	53 (5.0%)
TIS	55 (5.2%)
T1	419 (39.8%)
T2	279 (26.5%)
T3	57 (5.4%)
T4	27 (2.6%)
N stage (TNM)	
N0	425 (40.4%)
N1	205 (19.5%)
N2	77 (7.3%)
N3	58 (5.5%)
M stage (TNM)	
M0	501 (47.6%)
M1	139 (13.2%)
Ki-67 status	
Low/intermediate	238 (22.6%)
High	680 (64.6%)

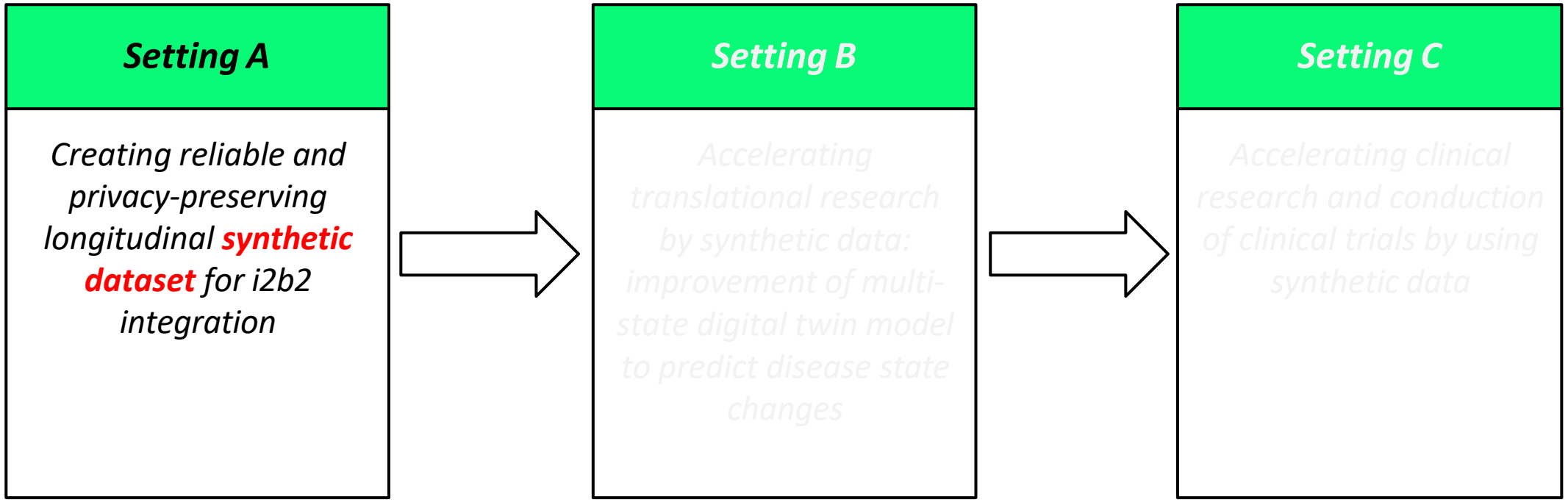
Study Population: patients distribution





Zambelli A. et al. jbres 2024

Results

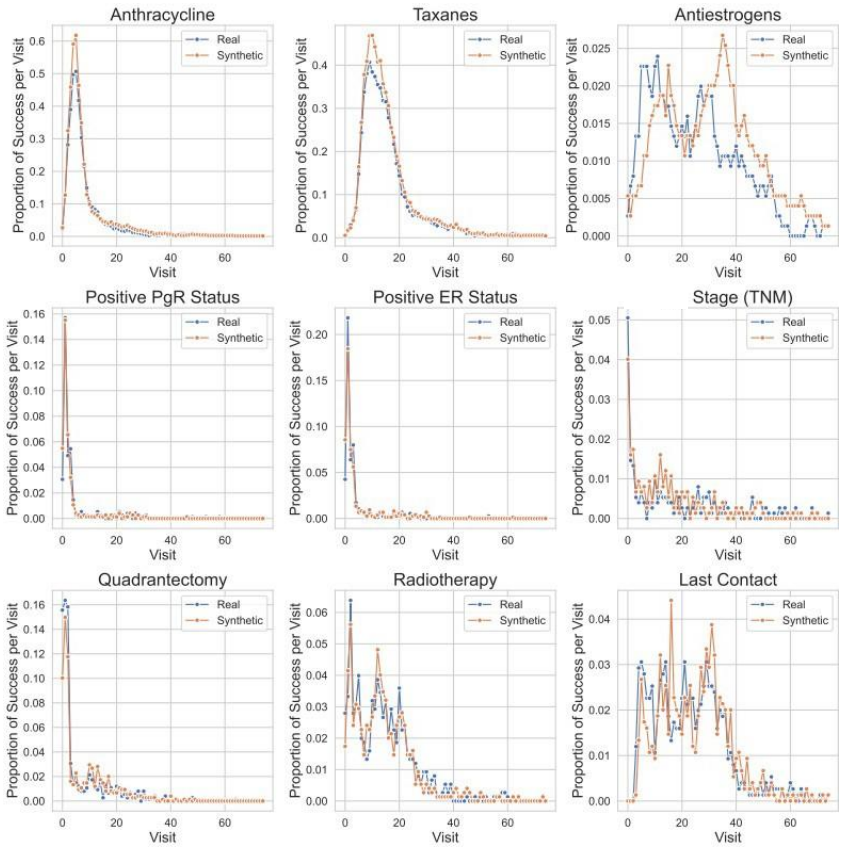

Setting A
Creating reliable and privacy-preserving longitudinal synthetic dataset for i2b2 integration
Setting B
Accelerating translational research by synthetic data: improvement of multi-state digital twin model to predict disease state changes
Setting C
Accelerating clinical research and conduction of clinical trials by using synthetic data

Synthetic Validation Framework

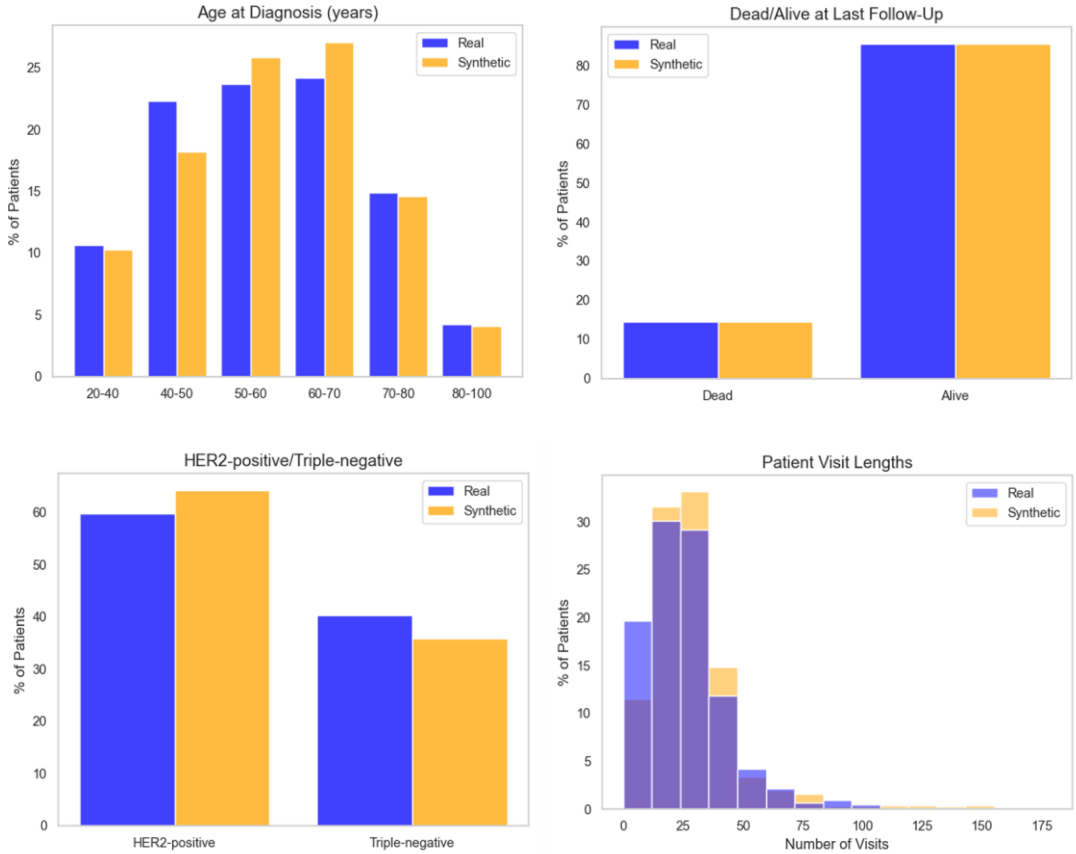
Fidelity	• Metrics: Dimension-wise probability test and TVD for discrete data, MMD and Kolmogorov-Smirnov test for continuous data. SMAPE, Granger casuality test and correlation for time series
Correlation	• Metrics: Pearson pairwise correlations
Privacy	• Ensure SD datasets do not replicate or closely resemble the training data
Utility	• Compares Kaplan-Meier survival curve for real vs SD (log-rank test)

Setting A (fidelity)

Time-dependent variables

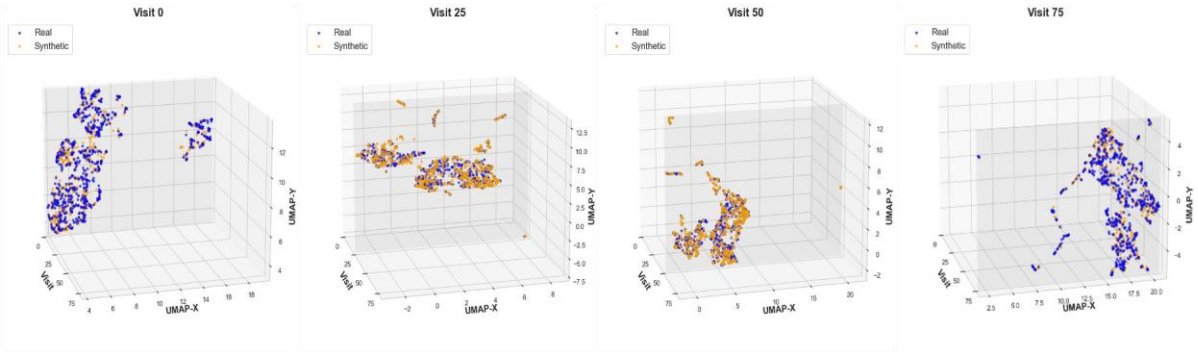


Non time-dependent variables

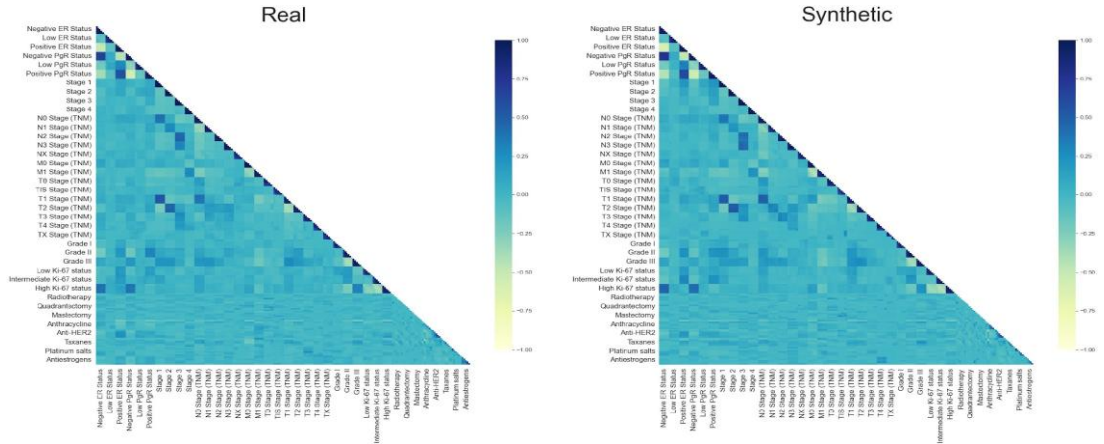


Setting A (fidelity and correlation)

- For a visual evaluation, **UMAP embeddings** were made per visit
- High overlap between real and synthetic embeddings across multiple timepoints ➔ similarity in data structure and temporal patterns



- **Pearson Correlation Matrices** show relationships between treatments, procedures, and clinical variables across visits
- SD preserves the correlation structure of real data, with strong consistency over time



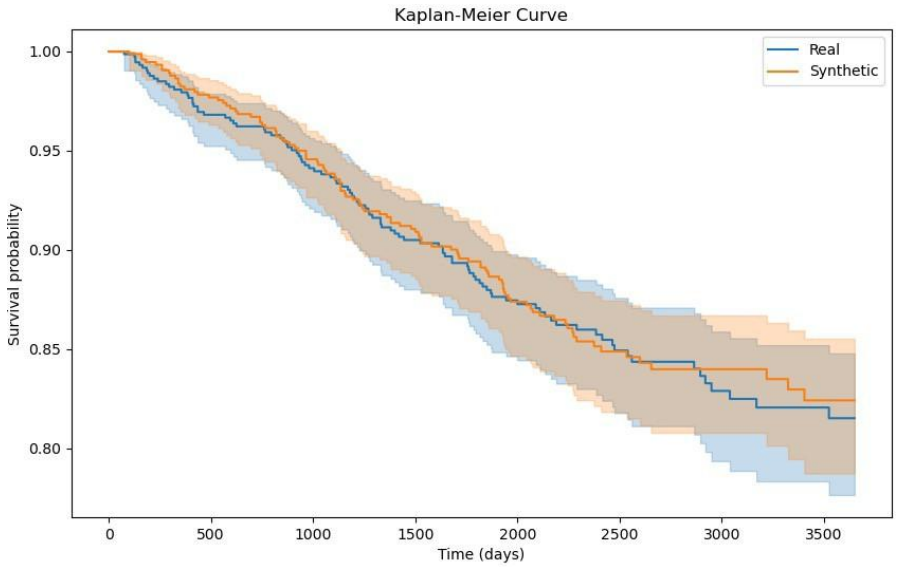
Setting A (privacy and utility)

- **Dataset attack** performed to test privacy
- Results for accuracy, precision and recall stay at 50% ➔ similar to random guessing
- SD cannot be linked back to specific real records ➔ strong privacy preservation

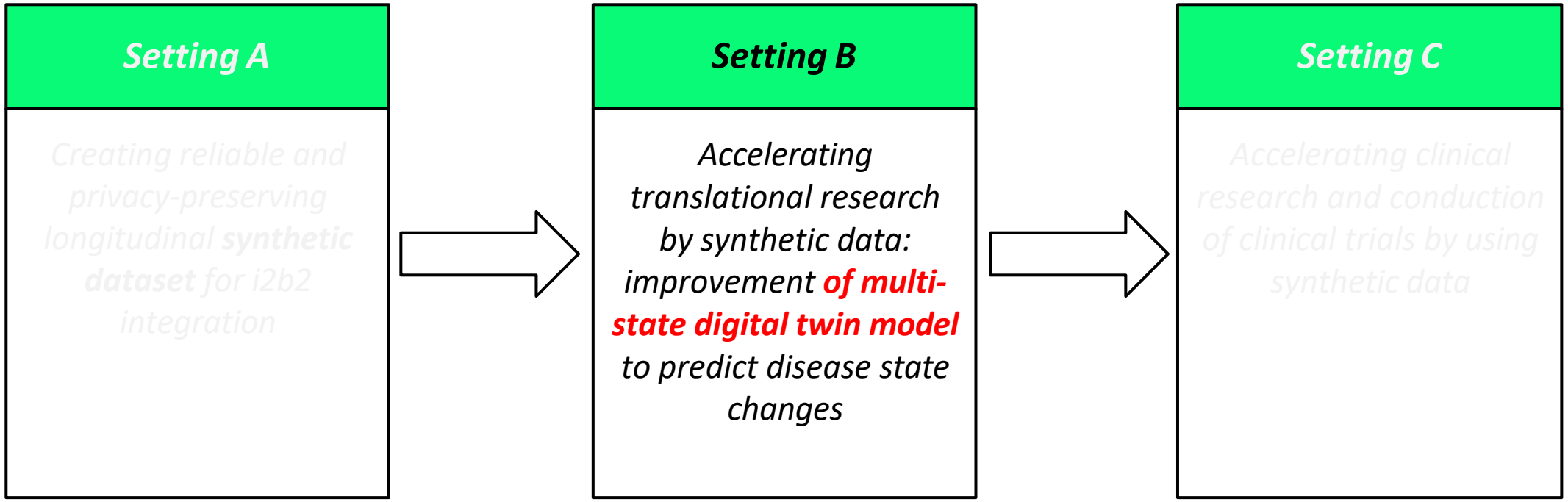
Privacy Dataset Attack

Accuracy	0.5
Precision	0.5
Recall	0.5

- Utility assessed using **Kaplan-Meier** survival curves
- A **log-rank test** obtained from comparing the survival curves of real and synthetic patients give a p-value of 0.773 ➔ no statistically significant difference in survival

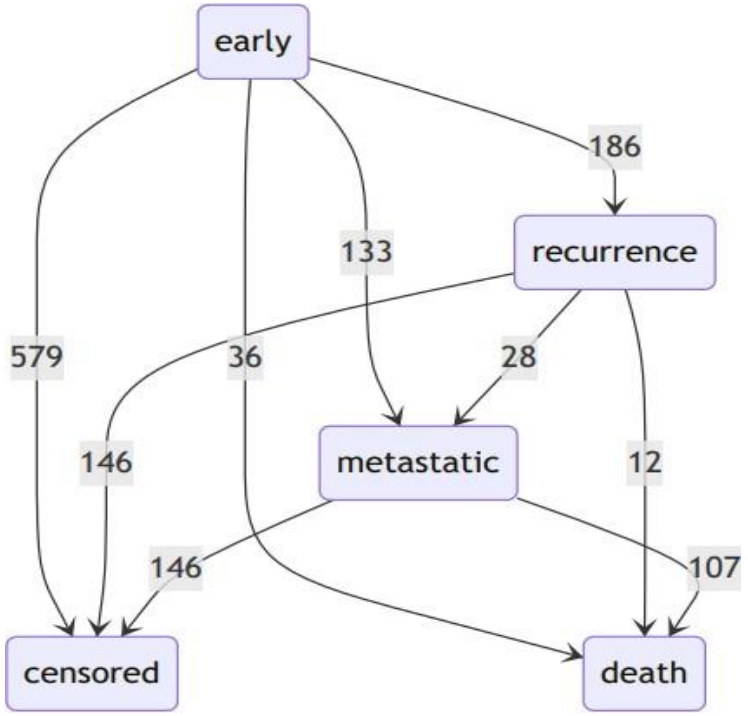


Results


Setting A
Creating reliable and privacy-preserving longitudinal synthetic dataset for i2b2 integration
Setting B
Accelerating translational research by synthetic data: improvement of multi-state digital twin model to predict disease state changes
Setting C
Accelerating clinical research and conduction of clinical trials by using synthetic data

Setting B

Disease State Transition Diagram



C-Index Performance

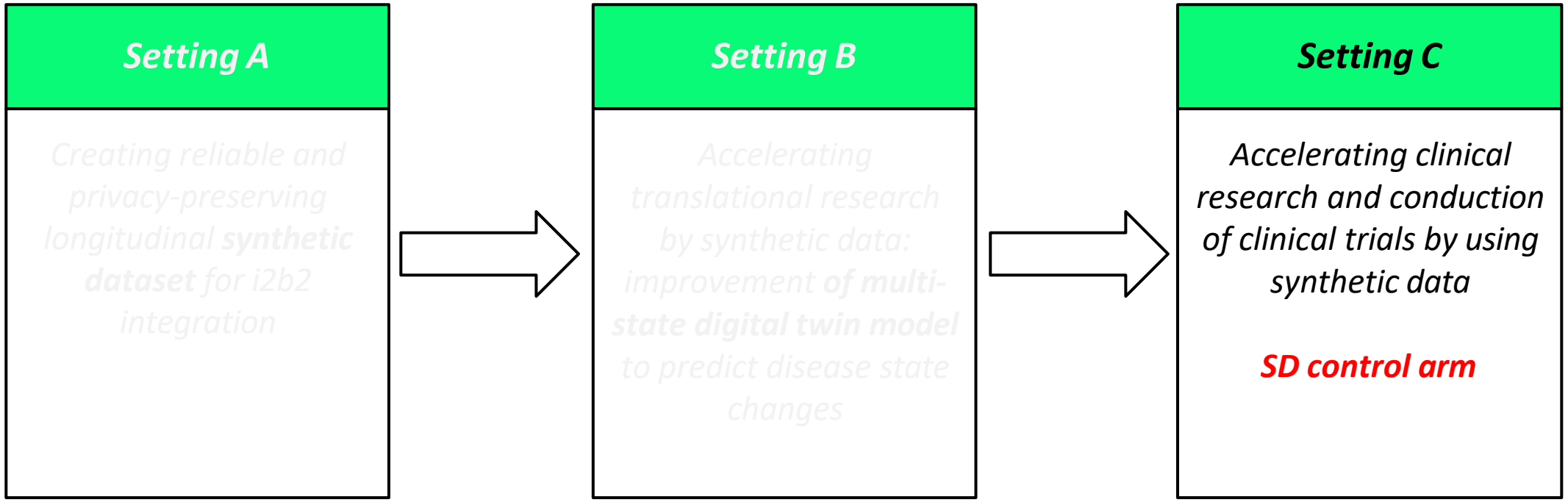
- Real samples (n)
- Synthetic samples (n, 3n, 2n plus n real samples)

Average C-Index	N samples	Real (n samples)	Synthetic (n samples)	Synthetic (3n samples)	Real+Synthetic (3n samples)
From Early	739	0.701	0.704	0.719	0.728
From Recurrence	146	0.778	0.766	0.835	0.858
From Metastatic	205	0.717	0.719	0.765	0.742

Average C-Index	N samples	Real (n samples)	Synthetic (n samples)	Synthetic (3n samples)	Real+Synthetic (3n samples)
From Early to Recurrence	146	0.756	0.725	0.749	0.737
From Early to Metastatic	105	0.765	0.694	0.771	0.756
From Early to Death	27	0.583	0.6941	0.636	0.691
From Recurrence to Metastatic	24	0.581	0.548	0.727	0.735
From Recurrence to Death	9	0.975	0.985	0.945	0.981
From Metastatic to Death	83	0.717	0.719	0.765	0.742

Augmented datasets improve C-index scores significantly

Results


Setting A
Creating reliable and privacy-preserving longitudinal synthetic dataset for i2b2 integration
Setting B
Accelerating translational research by synthetic data: improvement of multi-state digital twin model to predict disease state changes
Setting C
Accelerating clinical research and conduction of clinical trials by using synthetic data
SD control arm

Setting C

Real Dataset:

- 49 patients with HER2-positive, N0 BC, T ≤2 cm, treated with taxanes and trastuzumab in adjuvant phase
- 32 patients with HER2-positive, N0 BC, T ≤2 cm, treated with taxanes, trastuzumab and anthracyclines in adjuvant phase

Synthetic Dataset:

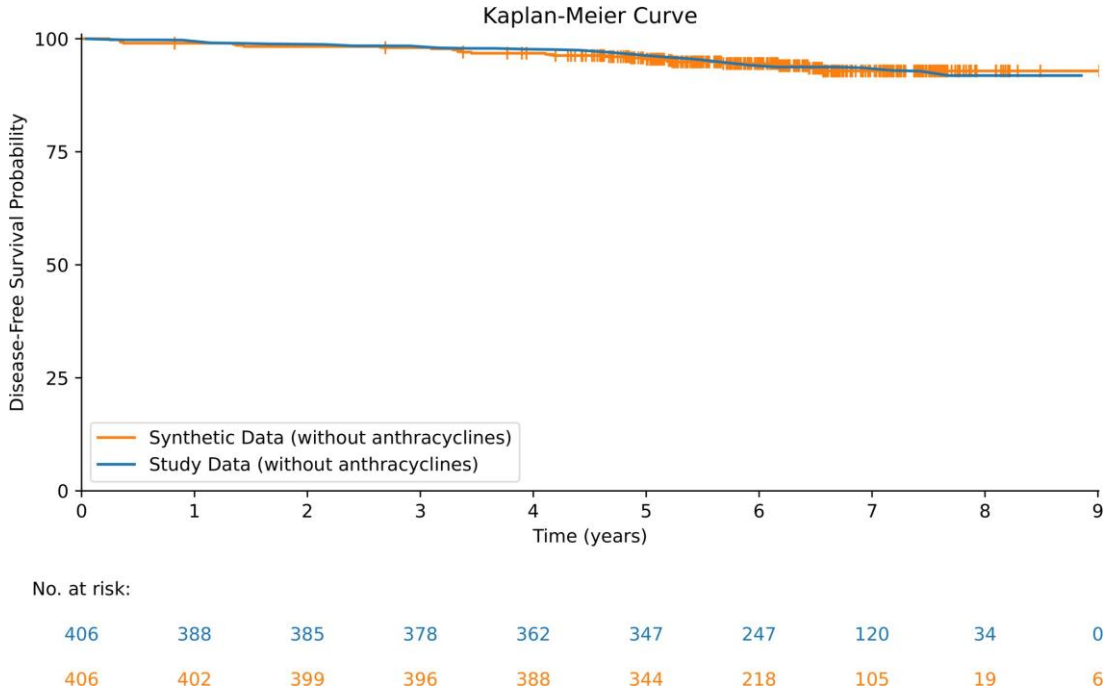
- Two synthetic cohorts of 406 patients to match the cohort size in the original study: one treated with taxane and trastuzumab and the other treated with taxane, trastuzumab and anthracyclines in the adjuvant phase

Tolaney, NEJM 2015; Tolaney, JCO 2019; Tolaney, Lancet 2023

APT trial: Study data vs. SD (w/o A)

Disease Free Survival:

- Kaplan-Meier curves show close alignment between synthetic and real study results
- SD reduced CI by ~72% ➔ greater statistical robustness vs the small real dataset (49 pts)
- Utility of SD in mimicking RW study outcomes



Event Comparison:

- Number of DFS events in real and synthetic
- The synthetic cohort accurately reflects the progression and survival outcomes observed in the real study arm

Time (years)	Point Est DFS (95% CI) Study (%)	No. Events Study	Point Est DFS (95% CI) Synthetic (%)	No. Events Synthetic
3	98.5 (97.2-99.7)	6	98.0 (96.1-99.0)	8
5	96.3 (94.4-98.2)	14	95.5 (92.9-97.1)	18
7	93.3 (90.4-96.2)	23	92.9 (89.3-95.3)	24

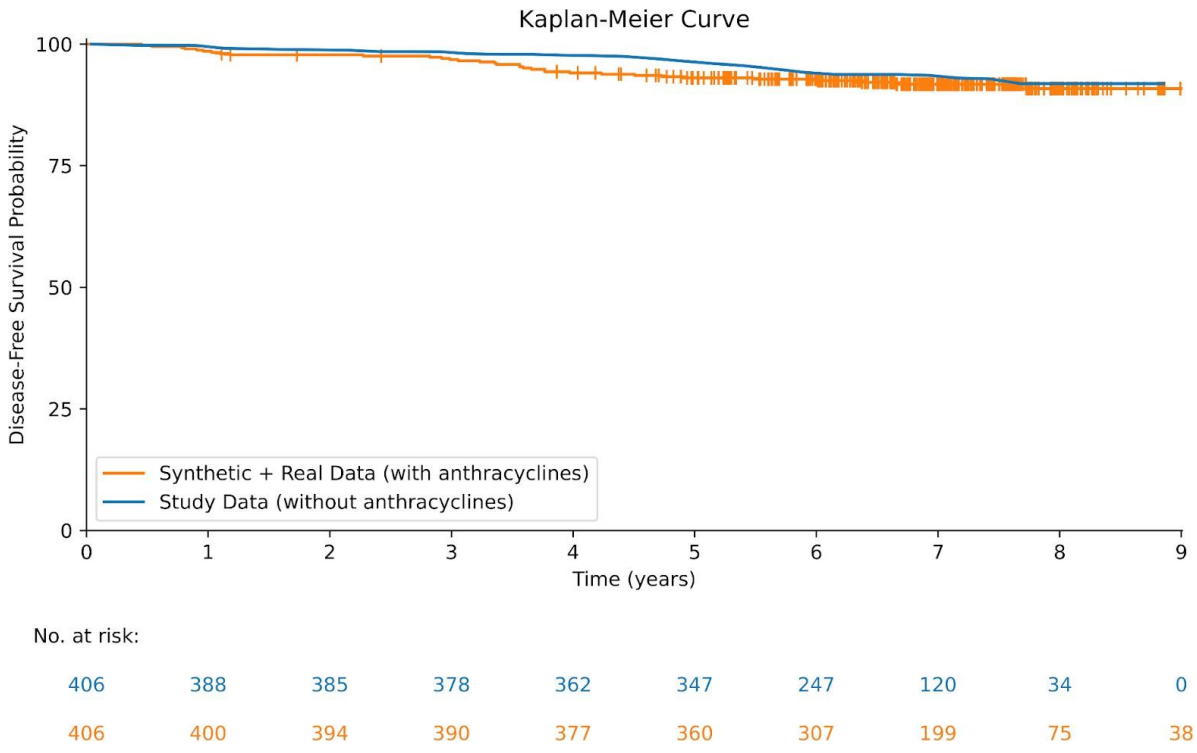
APT trial: Study data vs SD (with A)

Disease Free Survival:

- Comparability of Kaplan-Meier curves
- SD reduced CI by ~77% ➔ greater statistical robustness vs the small real dataset (32 pts)

Creation of Synthetic Control Arm:

- Strengthened evaluation of treatment efficacy for early-stage, node-negative, HER2-positive breast cancer
- Enhanced validity by addressing biases from the lack of a traditional control arm
- Reduced the need for enrolling real participants in the control group



Conclusion

- SD can be **generated from robust and accurate RW dataset**
- A **digital twin solution** for longitudinal multistage transition model can be implemented
- SD can replicate **clinical study** endpoints, enabling the creation of **artificial control arms**

Thank you

Azienda Ospedaliera
Papa Giovanni XXIII
Bergamo

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

azambelli@asst-pg23.it

alberto.zambelli@unimib.it