

A modern, single-story house with large glass windows is situated on a grassy hill. The house has a flat roof and a chimney. In the background, there are several tall, dark evergreen trees. The overall scene is misty or foggy, with a soft, diffused light. The text is overlaid on the upper right portion of the image.

Leveraging Synthetic Data in Precision Medicine: A Case Study from Our Experience

Maurizio Polano
mpolano@cro.it

Beyond Efficacy: Rethinking What “Working” Really Means in Medicine

Can it work? Does it work? Is it worth it?

The testing of healthcare interventions is evolving

The British pioneer clinical epidemiologist Archie Cochrane defined three concepts related to testing healthcare interventions.¹ Efficacy is the extent to which an intervention does more good than harm under ideal circumstances (“Can it work?”). Effectiveness assesses whether an intervention does more good than harm when provided under usual circumstances of healthcare practice (“Does it work in practice?”). Efficiency measures the effect of an intervention in relation to the resources it consumes (“Is it worth it?”). Trials of efficacy and effectiveness have also been described as explanatory and management trials, respectively,² and efficiency trials are more often called cost effectiveness or cost benefit studies.

Even if an intervention works astonishingly well in a “Can it work?” study, it may not work well in usual care. Effectiveness in the community depends not only on efficacy but also on diagnostic accuracy, provider compliance, patient adherence, and the coverage of health services.³ Misdiagnosis can result in the wrong people getting or not getting the treatment. Providers often fail to prescribe or administer the treatment properly. Patients typically take less than half of prescribed treatments. “High tech,” expensive, or new interventions are usually not available in all communities in the developed countries or to most communities in the rest of the world. To paraphrase Muir Gray, what works well at the Sloan Kettering (a high tech cancer

General practice
p 676

BMJ 1999;319:652–3

**Can it work?
Does it work?
Is it worth it?**

Rare Cancers Represent an Unmet Need That Cannot Be Resolved by Conventional Drug Development Paradigms

Challenges to developing drugs to treat rare cancers are numerous:

- (1) extreme difficulty enrolling sufficient number of patients to clinical trials, which approaches being almost impossible for ultra-rare cancers;
- (2) decreased financial incentives for drug development;
- (3) inadequate research into natural history and cancer biology;
- (4) virtually unmanageable challenges to conduct randomized trials because of small patient numbers and lack of an appropriate standard of care for the control arms.

FREE ACCESS | POLICY AND PRACTICE | June 11, 2025



Imperative of Comprehensive Molecular Profiling as Standard of Care for Patients With Rare Cancers

Authors: [Vivek Subbiah, MD](#) , and [Razelle Kurzrock, MD](#)  | [AUTHORS INFO & AFFILIATIONS](#)

Publication: JCO Oncology Practice • [Newest Articles](#) • <https://doi.org/10.1200/OP-25-00064>

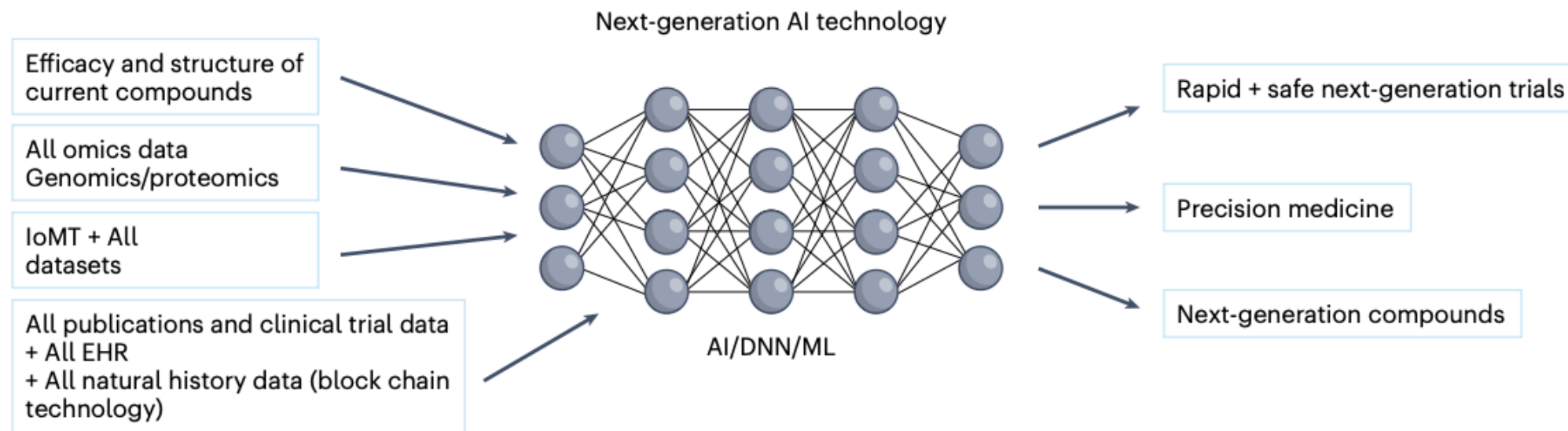
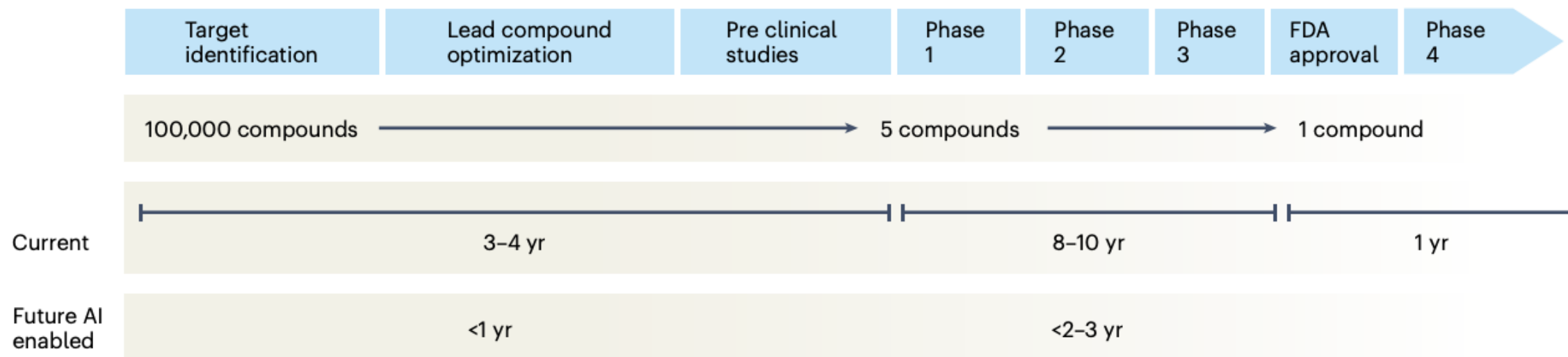
1,060  4



In the current evolving landscape of cancer treatment, the notion that all patients with advanced cancer, but especially those with rare cancers, must receive comprehensive molecular profiling as part of their care is becoming increasingly compelling.¹ In a viewpoint, we posited that universal tumor genomic testing, in particular next-generation sequencing



How do clinical trials progress?



Subbiah, V. et.al (2023)

<https://doi.org/10.1038/s41591-022-02160-z>

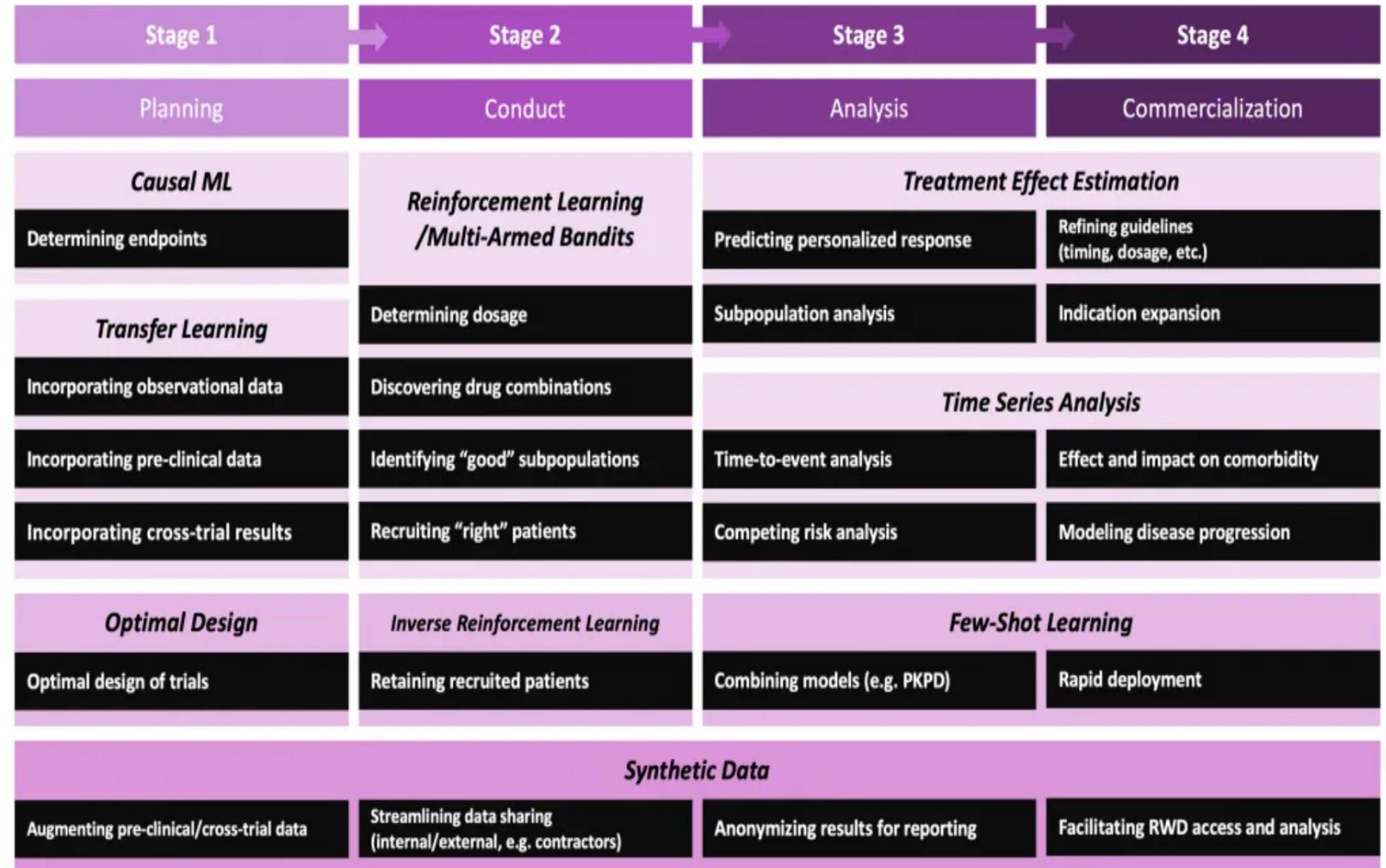
THE CHALLENGES OF CONDUCTING AN Randomized controlled trials (RCT)

(Stage 1) **planning** of a clinical trial that targets the new treatment,

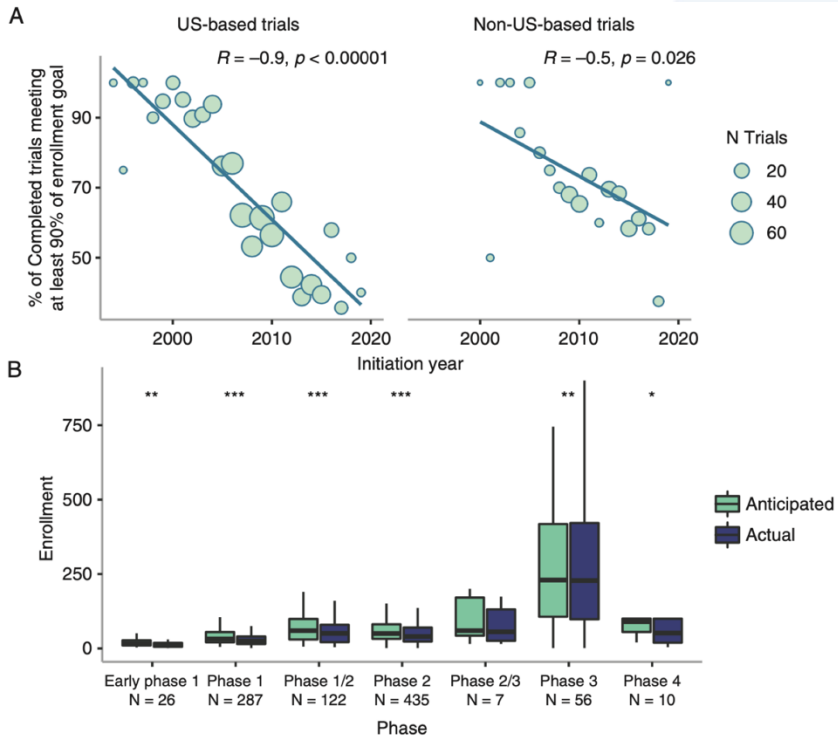
(Stage 2) **conduct** of the planned trial,

(Stage 3) **analysis** of the results obtained from the trial,

(Stage 4) **clinical-use** of the treatment if the trial has been successful.



Clinical Trials fails



37.9% failed to reach their enrollment (all phases, less in phase 2/3)

Neuro-Oncology

XX(XX), 1–14, 2023 | <https://doi.org/10.1093/neuonc/noad036> | Advance Access date 9 February 2023

A critical analysis of neuro-oncology clinical trials

Yeonju Kim, Terri S. Armstrong[✉], Mark R. Gilbert, and Orieta Celiku[✉]

Neuro-Oncology
ADVANCES

OXFORD
UNIVERSITY PRESS

[Neurooncol Adv.](#) 2024 Jan-Dec; 6(1): vdad169.

Published online 2024 Jan 10. doi: [10.1093/noajnl/vdad169](https://doi.org/10.1093/noajnl/vdad169)

PMCID: PMC10838133

PMID: [38312230](https://pubmed.ncbi.nlm.nih.gov/38312230/)

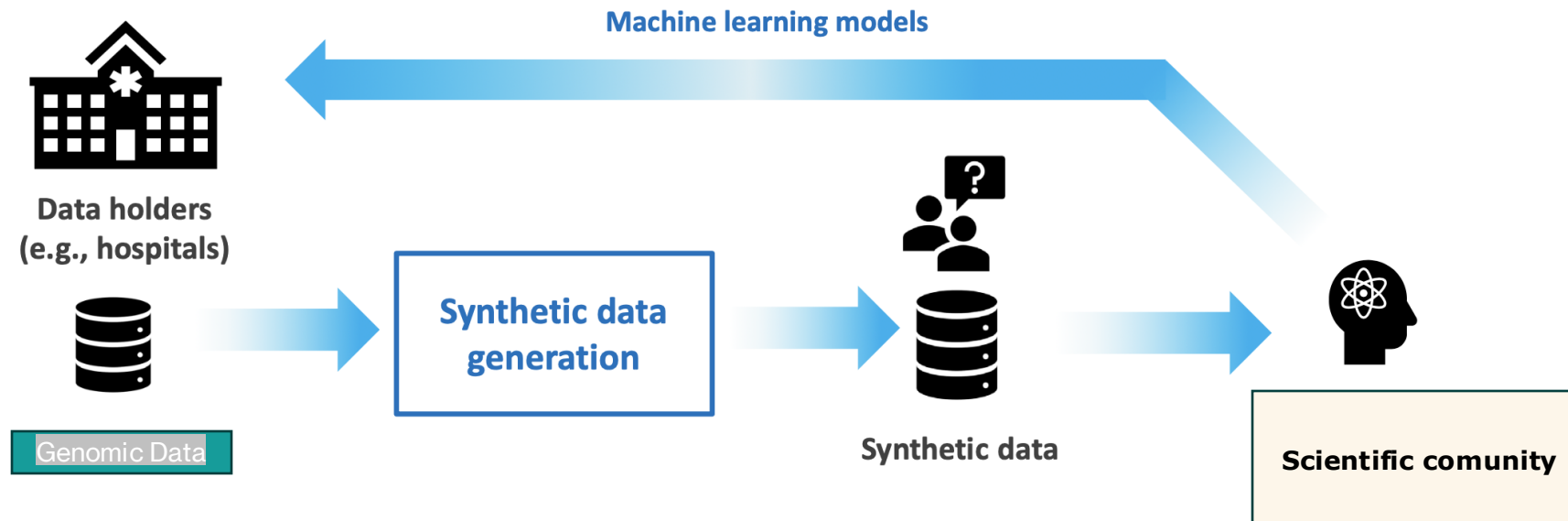
Adult neuro-oncology trials in the United States over 5 decades: Analysis of trials completion rate to guide the path forward

Key Points

- Neuro-oncology clinical trial completion rates have significantly decreased over the past 2 decades, from 78% to 64%.
- Fifty percent of the US population affected by neuro-oncology diseases have limited access to neuro-oncology trials.

Synthetic Data to accelerate research in pharma-oncology

Synthetic data are artificial data generated by an algorithm trained to learn all the essential characteristics of a real dataset. The new data are neither a copy nor a representation of the real data. Since they are not real data; they are not regulated by particular limitations so they can be easily accessed and shared



Synthetic Data to accelerate research in pharma-oncology



[Login](#) [Register](#) [Need Help?](#)

[ABOUT](#) [DISCOVERY](#) [SUBMISSION](#) [ACCESS](#)

Synthetic Data

One of the limitations in genomics research is that human genomics data is not openly available; access must be controlled according to participant consent agreements and data protection regulations such as GDPR. Obtaining authorization to access such data can sometimes take a long time, resulting in delays to important research work. In this context, synthetic genomic and phenotype data can be useful resources for researchers to avoid these delays.

Synthetic data are artificially generated datasets, often created with algorithms, which can be used without the need for authorization to test new products and tools, build technical demonstrators, validate data models, and train AI models. The EGA provides access to synthetic cohort datasets augmented with rich synthetic metadata that overcomes these real data usage restrictions. Whilst synthetic datasets are not included in the general EGA mandate and services, we can consider such submissions and evaluate their acceptance on the basis of their unique use cases not already covered by existing synthetic datasets. Access to synthetic data studies is managed by the EGA Helpdesk Data Access Committee.

Study ID	Title	Located in
EGAS00001002472	CINECA synthetic cohort EUROPE UK1 referencing fake samples	Central EGA
EGAS00001005591	Synthetic data - Genome in a Bottle	Central EGA
EGAS00001005042	Test Study for EGA using data from 1000 Genomes Project - Phase 3	Central EGA
EGAS00001005702	Human genomic and phenotypic synthetic data for the study of rare diseases	Central EGA
EGAS50000000190	EOSC4Cancer Synthetic Colorectal Cancer Genomic data	Central EGA
EGAS50000000086	Synthetic - FEGA Sweden Heilsa synthetic dataset December 2023	Federated EGA Sweden
EGAS500000000678	Synthetic - GDI synthetic data	Federated EGA Spain

The WHO guidance on Ethics & Governance of Artificial Intelligence for Health

28 June 2021 | Guideline



[Download \(1.9 MB\)](#)

- 1) **Transparency of models:** interpretability and explainability
- 2) **Reliability of models:** independent validation of generated AI-models
- 3) **Protection of data and data sharing:** compliance with GDPR (EU)

Synthetic ≠ Fake: Reimagining Data for Science and Privacy

Synthetic data could be better than real data

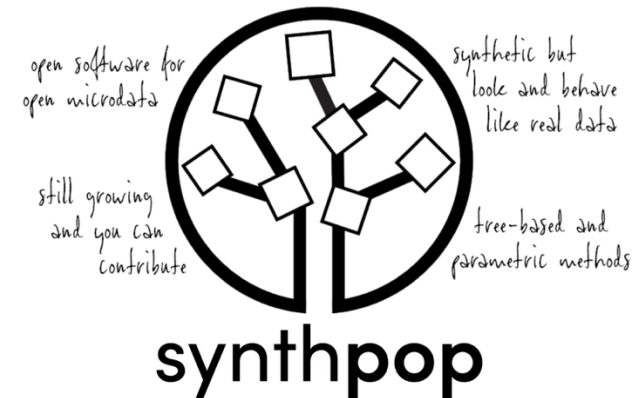
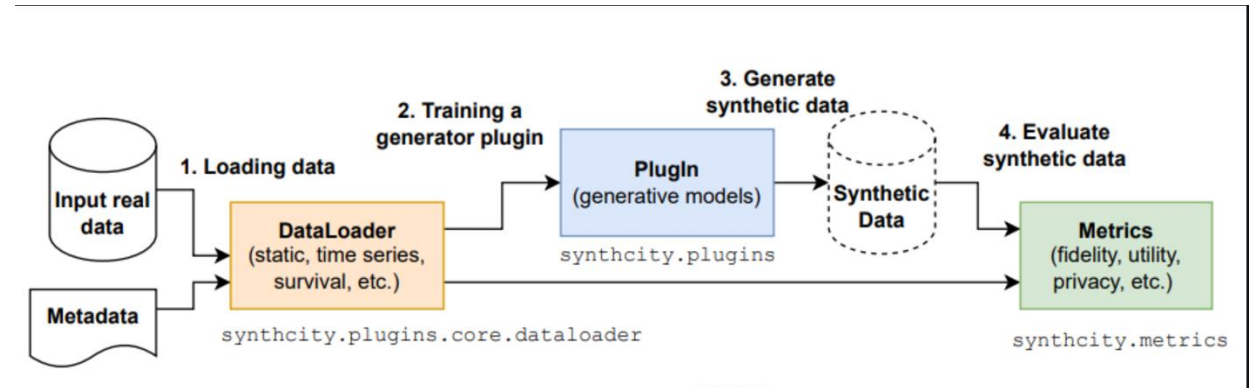
Machine-generated data sets have the potential to improve privacy and representation in artificial intelligence, if researchers can find the right balance between accuracy and fakery.

By [Neil Savage](#)



Credit: Janelle Barone

<https://www.nature.com/articles/d41586-023-01445-8>

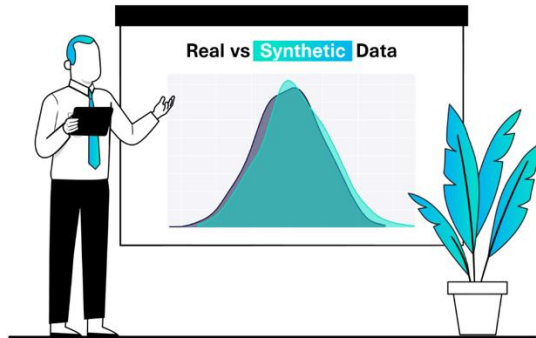


From Performance to Privacy: The Multidimensional Assessment of Synthetic Data

SDMetrics



Synthetic Data Metrics (SDMetrics) is an [open source](#) Python library for evaluating tabular synthetic data. Compare synthetic data against real data using a variety of metrics, generate visual reports and share them with your team.



<https://docs.sdv.dev/sdmetrics>

Synthetic Data Privacy Metrics

Amy Steier Gretel.ai amy@gretel.ai	Lipika Ramaswamy Gretel.ai lipika@gretel.ai	Andre Manoel Gretel.ai andre.manoel@gretel.ai	Alexa Haushalter Gretel.ai alexa@gretel.ai
---	--	--	---

Abstract

Recent advancements in generative AI have made it possible to create synthetic datasets that can be as accurate as real-world data for training AI models, powering statistical insights, and fostering collaboration with sensitive datasets while offering strong privacy guarantees. Effectively measuring the empirical privacy of synthetic data is an important step in the process. However, while there is a multitude of new privacy metrics being published every day, there currently is no standardization. In this paper, we review the pros and cons of popular metrics that include simulations of adversarial attacks. We also review current best practices for amending generative models to enhance the privacy of the data they create (e.g. differential privacy).

SPRINGER NATURE Link

[Find a journal](#) [Publish with us](#) [Track your research](#) [Search](#)

[Home](#) > [Digital Society](#) > [Article](#)

Unraveling the Regimes of Synthetic Data Metrics: Expectations, Ethics, and Politics

Brief Communication | [Open access](#) | Published: 04 June 2025

Volume 4, article number 44, (2025) [Cite this article](#)

Trust, Test, Validate: Metrics for Synthetic Data

- Sanity checks

Metric	Description
data_mismatch	Average number of columns with datatype(object, real, int) mismatch between real and synthetic data.
common_rows_proportion	The proportion of rows in the real dataset leaked in the synthetic dataset.
nearest_syn_neighbor_distance	Average distance from the real data to the closest neighbor in the synthetic dataset.
close_values_probability	The probability of close values between the real and synthetic data.
distant_values_probability	Average distance from the real data to the closest neighbor in the synthetic dataset.

- Synthetic Data quality

Metric	Description
performance.xgb	Train an XGBoost classifier/regressor/survival model on real data(gt) and test on synthetic data.
performance.linear	Train a Linear classifier/regressor/survival model on real data(gt) and test on synthetic data.
performance.mlp	Train a Neural Net classifier/regressor/survival model on the real data(gt) and test on synthetic data.
performance.featurization_distance	Train a model on the synthetic data and a model on the real data. Compute the distance between the two models.
detection.gmm	Train a GaussianMixture model to differentiate the synthetic data from the real data.
detection.xgb	Train an XGBoost model to differentiate the synthetic data from the real data.
detection.mlp	Train a Neural net to differentiate the synthetic data from the real data.
detection.linear	Train a Linear model to differentiate the synthetic data from the real data.

- Statistical tests

Metric	Description
inverse_kl_divergence	The average inverse of the Kullback-Leibler Divergence.
ks_test	The Kolmogorov-Smirnov test.
chi_squared_test	The p-value. A small value indicates that we can reject the null hypothesis and that the data is not from the same distribution.
max_mean_discrepancy	Empirical maximum mean discrepancy.
jensenshannon_dist	The Jensen-Shannon distance (metric) between two probability arrays. This is the square root of the Jensen-Shannon divergence.
wasserstein_dist	Wasserstein Distance is a measure of the distance between two probability distributions.
prdc	Computes precision, recall, density, and coverage given two manifolds.
alpha_precision	Evaluate the alpha-precision, beta-recall, and authenticity scores.
survival_km_distance	The distance between two Kaplan-Meier plots(survival analysis).
fid	The Frechet Inception Distance (FID) calculates the distance between two distributions.

- Privacy metrics

Quasi-identifiers : pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier.

Metric	Description
k_anonymization	The minimum value k which satisfies the k-anonymity rule: each record has at least k other records with the same quasi-identifiers.
l_diversity	The minimum value l which satisfies the l-diversity rule: every generalization of a quasi-identifier must contain at least l distinct values for the sensitive attribute.
kmap	The minimum value k which satisfies the k-map rule: every combination of quasi-identifiers must map to at least k distinct values for the sensitive attribute.
delta_presence	The maximum re-identification risk for the real dataset from the synthetic dataset.
identifiability_score	The re-identification score on the real dataset from the synthetic dataset.
sensitive_data_reidentification_xgb	Sensitive data prediction from the quasi-identifiers using an XGBoost model.
sensitive_data_reidentification_mlp	Sensitive data prediction from the quasi-identifiers using a Neural Network model.



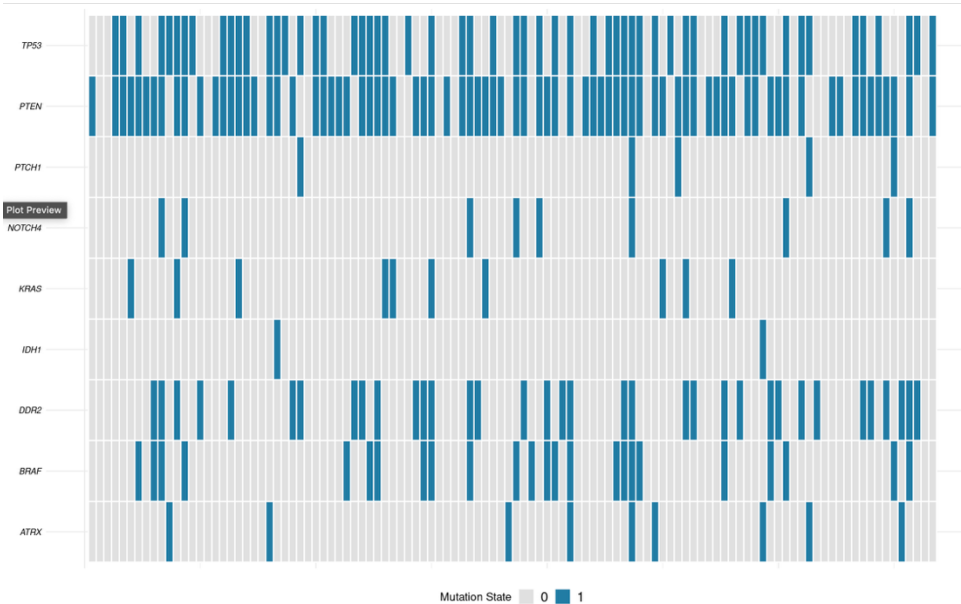
Tests

Are Synthetic Data Real Enough? Sanity, Statistics, and Privacy

1. SANITY CHECKS

Queste metriche misurano se i dati sintetici sembrano "finti".

Metrica	Interpreta così
<code>data_mismatch.score</code> = 0	✅ i domini dei valori sono compatibili
<code>common_rows_proportion</code> = 0.011	🔒 pochi duplicati dei dati reali → buono
<code>nearest_syn_neighbor_distance</code>	più alto = più distanti → <code>dummy_sampler</code> più vicino ai reali (possibil overfitting)
<code>close_values_probability.score</code>	🔍 più alto = valori troppo simili ai reali (rischio overfit)
<code>distant_values_probability.score</code>	🔍 più basso = dati meno distanti → plausibile



2. STATISTICAL METRICS

Confronta le distribuzioni statistiche reali vs sintetiche

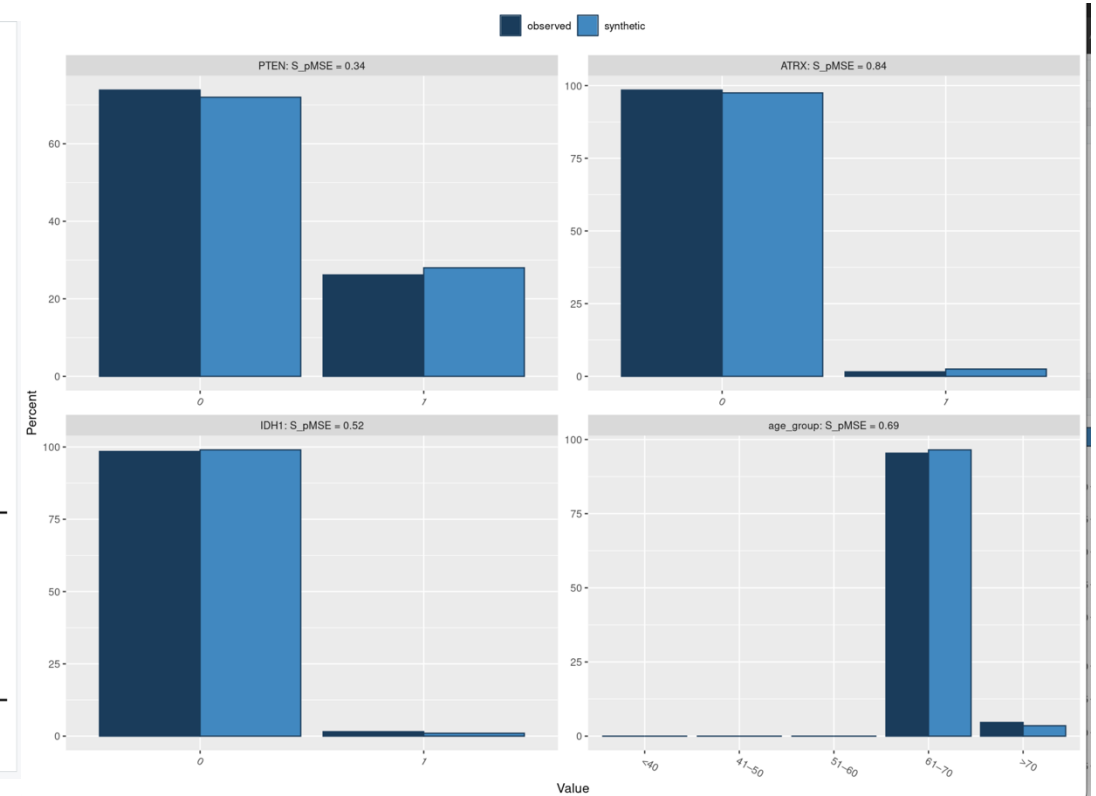
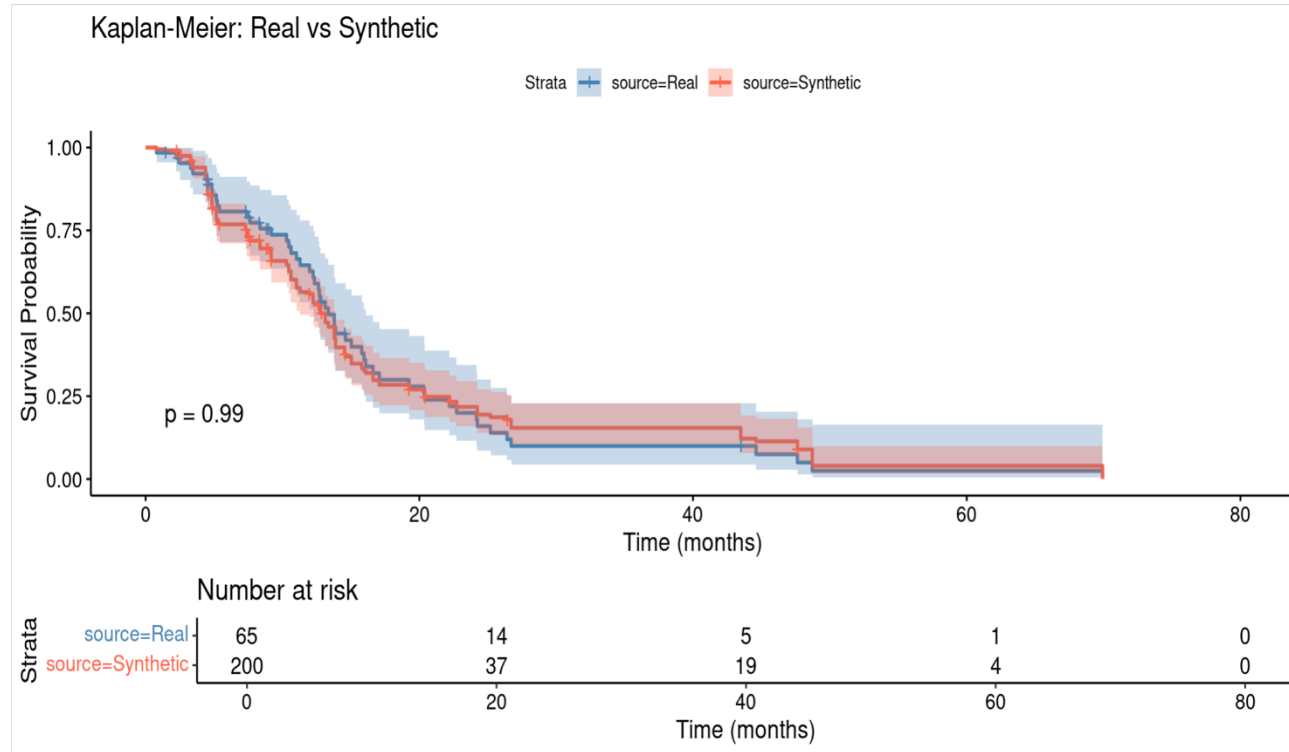
Metrica	<code>dummy_sampler</code>
<code>jsd.marginal</code> (0.012)	✅ Molto simile ai reali
<code>chi_squared_test.marginal</code> (0.75)	✅ Alta similarità
<code>ks_test.marginal</code> (0.946)	✅ Simile univariatamente
<code>mmd.joint</code> (0.043)	✅ Vicino ai reali anche in modo multivariato
<code>wasserstein_dist.joint</code> (0.214)	✅ Buono
<code>prdc.precision/recall/density/coverage</code>	✅ Ottimi valori (quasi 1)

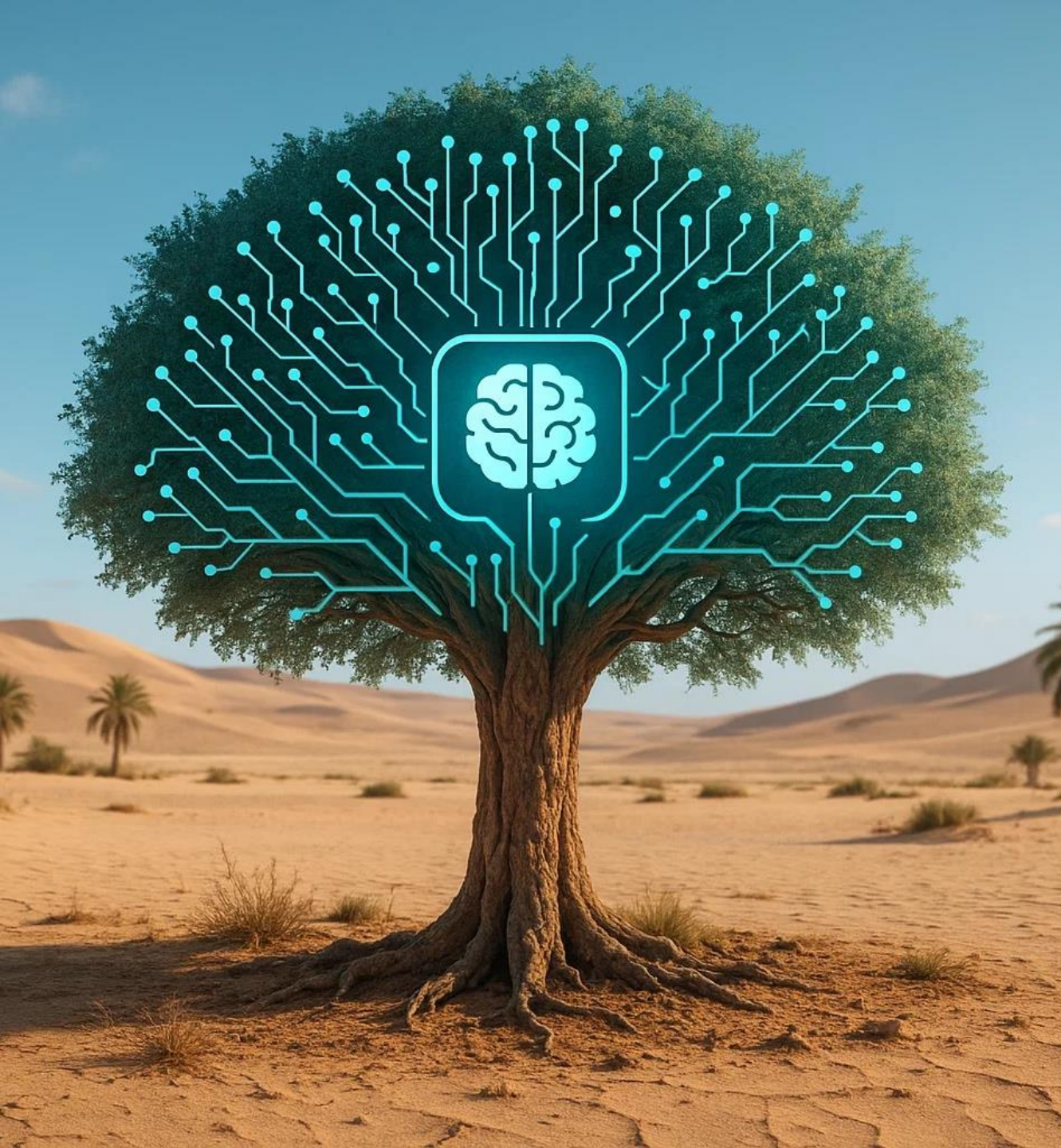
● In sintesi: `dummy_sampler` sta generando dati che sembrano **statisticamente simili ai reali**.

3. PRIVACY METRICS

Metrica	<code>dummy_sampler</code>	Interpretazione
<code>delta-presence</code> = 1.17	🔒 Buono (basso rischio)	
<code>k-anonymity.syn</code> = 11.5	🔒 Buona protezione	
<code>k-map</code> = 8.0	🔒 Accettabile	
<code>l-diversity.syn</code> = 2.0	🟡 Minima diversità	
<code>identifiability_score</code> = 0.5	⚠️ Più alto = più rischioso	
<code>DomiasMIA_BNAF.aucroc</code> = 0.5	✅ No signal (chance level)	
<code>DomiasMIA_prior.aucroc</code> = 0.654	⚠️ Qualche rischio, ma non eccessivo	

Are Synthetic Data Real Enough?



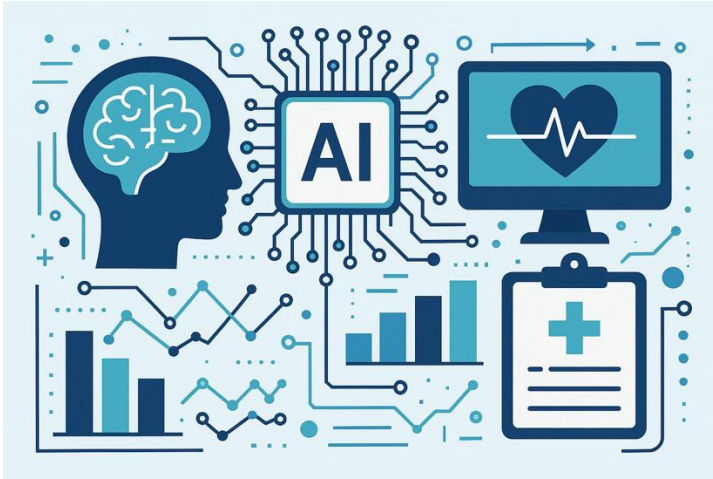


Synthetic data

The emerging risk of “model collapse” presents a challenge in maintaining the truthfulness and utility of synthetic data

This term refers to the [degradation of AI models](#) that occurs when they are repeatedly trained on data produced by other models. This recursive training can result in the perpetuation of models ‘[hallucinating](#),’ or overemphasising common patterns while underrepresenting rarer ones, leading to unrealistic synthetic data generation

How can Synthetic improve our work



Generate new evidence

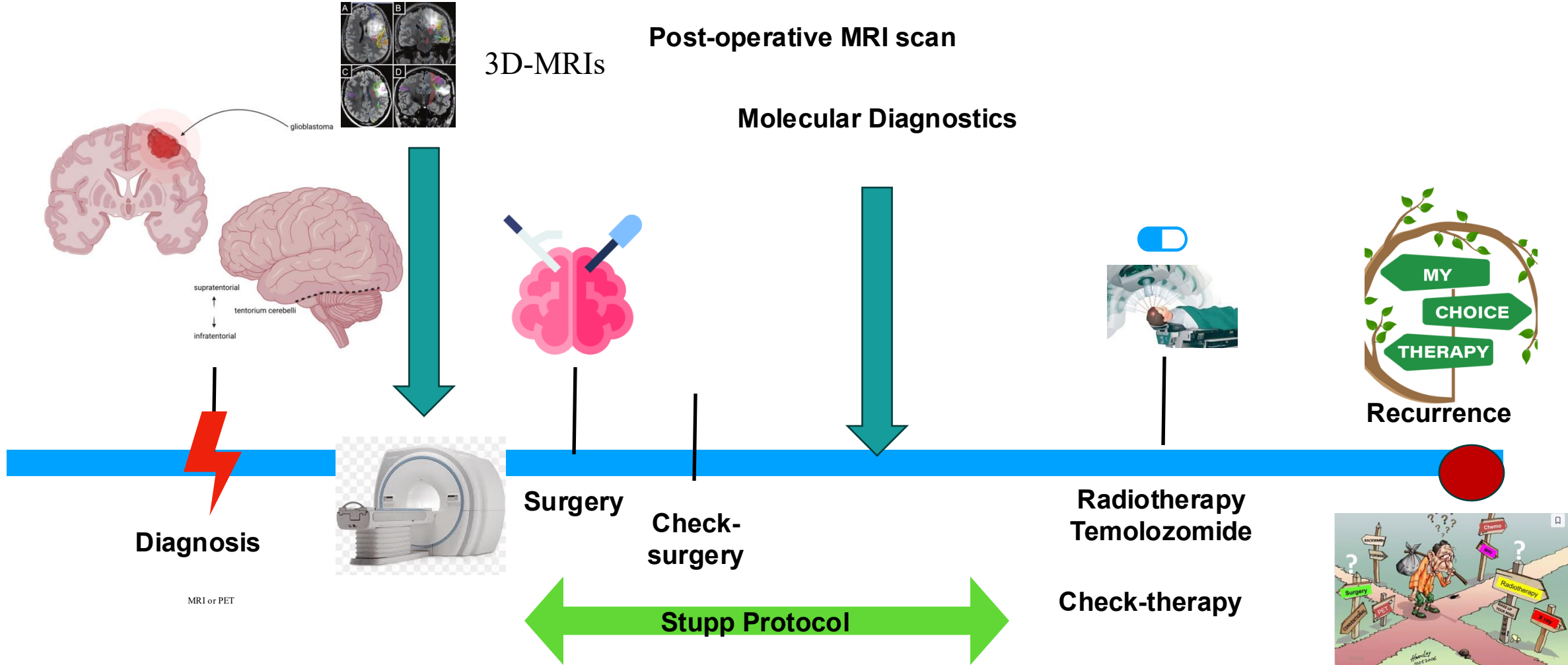
Synthetic Arm

Data augmentation

Data Sharing/Reproducibility (Privacy/GDPR)

The patient journey

Generate new evidence



GLIADEL (CW) SEEMS TO SUSTAIN A BETTER SURVIVAL

Generate new evidence

STATISTICAL APPROACH

UNIVARIATE

Characteristic	HR (95% CI) ¹	p-value
gliadel		
0	—	
1	0.53 (0.34 to 0.81)	0.003

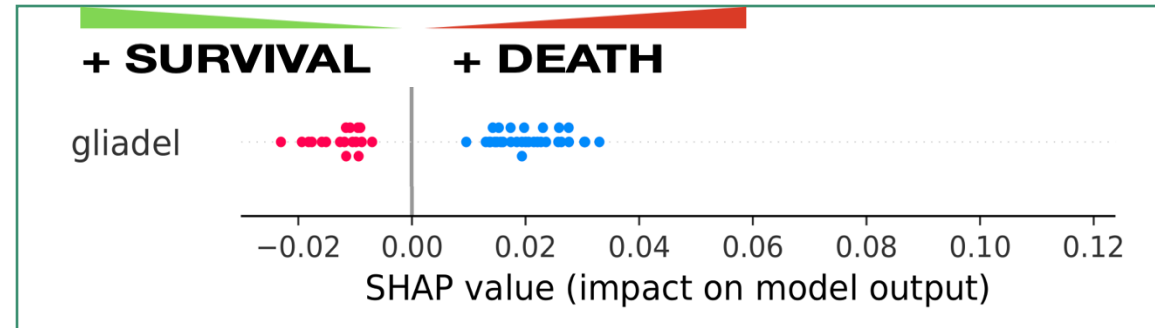
¹HR = Hazard Ratio, CI = Confidence Interval

MULTIVARIATE

Characteristic	HR (95% CI) ¹	p-value
gliadel		
0	—	
1	0.62 (0.40 to 0.96)	0.032
volumepre	1.01 (1.00 to 1.02)	0.25
EOR	0.94 (0.91 to 0.98)	0.002
mgmt_patho		
0	—	
1	0.47 (0.30 to 0.74)	<0.001

¹HR = Hazard Ratio, CI = Confidence Interval

MACHINE LEARNING APPROACH

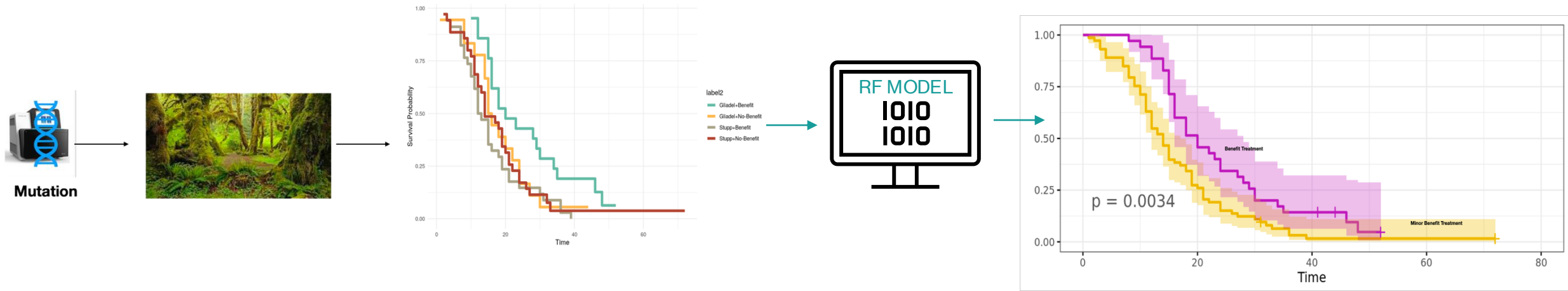


Shap dependence plot show how the model output varies by feature value.

When patients are treated with CW, we improve survival time

Patients most likely to benefit from a particular treatment can be identified using machine learning?

Generate new evidence



The objective is to identify the molecular status associated with full treatment benefit.

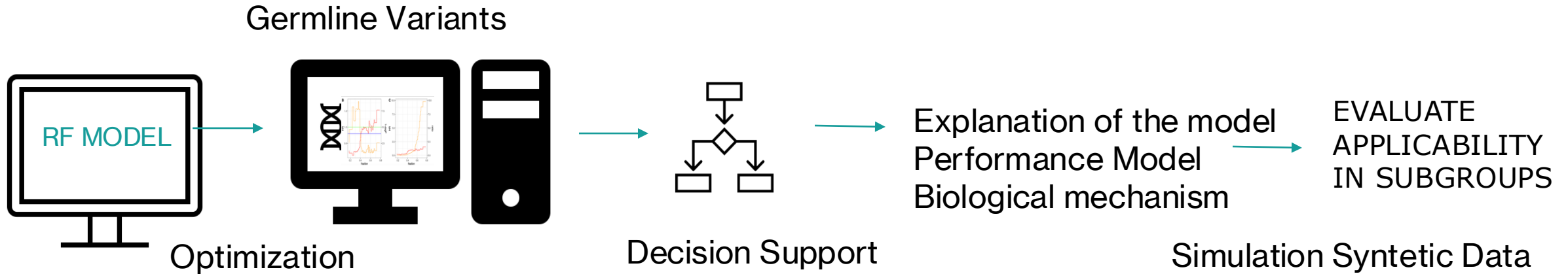
This finding lends support to the safety and efficacy of the treatment.

The result may provide a **rationale** for the creation of **a validation clinical trial** to evaluate the treatment.

1. RWD can be used to identify eligible patients
2. identify which subgroup respond better to treatment (develop other RCT)
3. use of generative procedure to validate model and understand the patients

DEVELOP THE FINAL MODEL

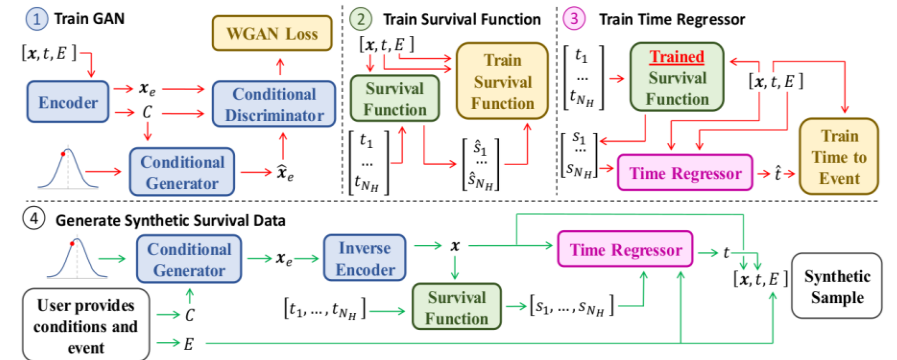
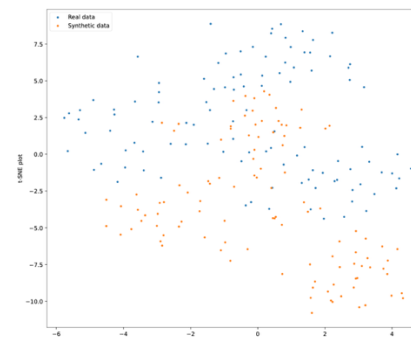
Generate new evidence



Generate **fair** data based on **unfair** data

- **Augment** small-sample data sets
- **Adapt** data to new domains
- **Simulate** unseen scenarios & **realistic** futures

Generate **realistic synthetic test sets** for ML model testing (3S)



Synthetic Control Arms for Rare Tumors:

A Realistic Alternative to Placebo

Generate Synthetic Arm

Comment

<https://doi.org/10.1038/s41591-023-02578-7>

Rethinking placebos: embracing synthetic control arms in clinical trials for rare tumors

César Serrano, Sara Rothschild, Guillermo Villacampa, Michael C. Heinrich, Suzanne George, Jean-Yves Blay, Jason K. Sicklick, Gary K. Schwartz, Sameer Rastogi, Robin L. Jones, Piotr Rutkowski, Neeta Somaiah, Víctor Navarro, Denise Evans & Jonathan C. Trent

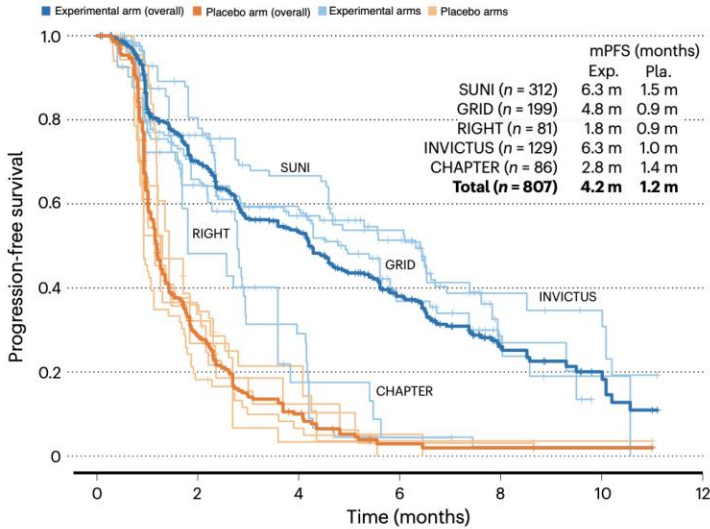
Check for updates

External comparator arms should be used when investigating novel therapies for gastrointestinal stromal tumor and other rare tumors to facilitate drug testing and regulatory approvals.

Rare cancers, although individually uncommon, account for approximately one-quarter of all malignancies. Population-based studies have

care that could be used as a comparator. Several measures were taken to maximize the likelihood of exposure to the experimental treatment arm, such as uneven randomizations, crossover designs at early response assessment. However, late-stage GIST has increased aggressiveness and tumor bulk after numerous lines of treatment and therefore the use of a placebo poses a substantial risk to patients. This was evidenced recently in the INVICTUS trial, which compared ripretinib to placebo after progression following three or more lines of treatment for metastatic disease¹. In this study, one-third of patients originally allocated to placebo during the double-blind period were

a Progression-free survival



b Overall survival

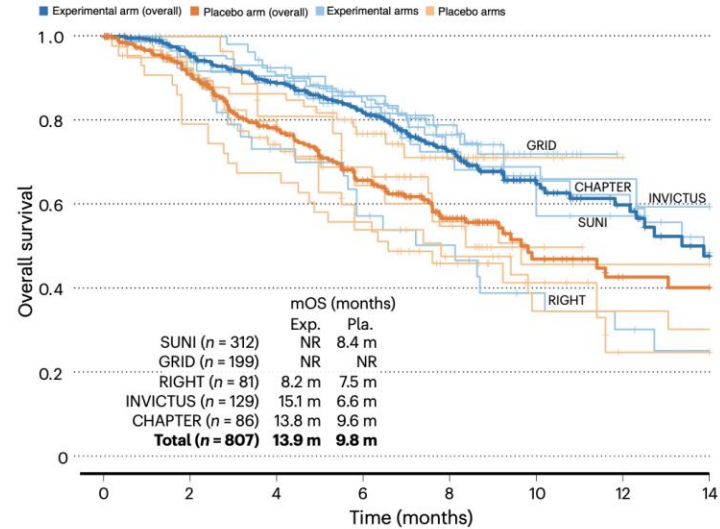


Fig. 1 | Progression-free and overall survival in GIST. a, b, Progression-free survival (a) and overall survival (b) Kaplan–Meier curves for the placebo and experimental arms in the five placebo-controlled RCTs of treatments for imatinib-resistant GIST. SUNI*, sunitinib versus placebo (asterisk denotes the

use of time to treatment failure rather than progression-free survival as primary endpoint); GRID, regorafenib versus placebo; RIGHT, imatinib versus placebo; INVICTUS, ripretinib versus placebo; CHAPTER, pimitespid versus placebo; mPFS, median progression-free survival; m, months.

No More Placebos? Rethinking Control Arms with Synthetic Data

Generate Synthetic Arm

<https://doi.org/10.1038/s41591-023-02488-0>

A synthetic control arm for refractory metastatic colorectal cancer: the no placebo initiative

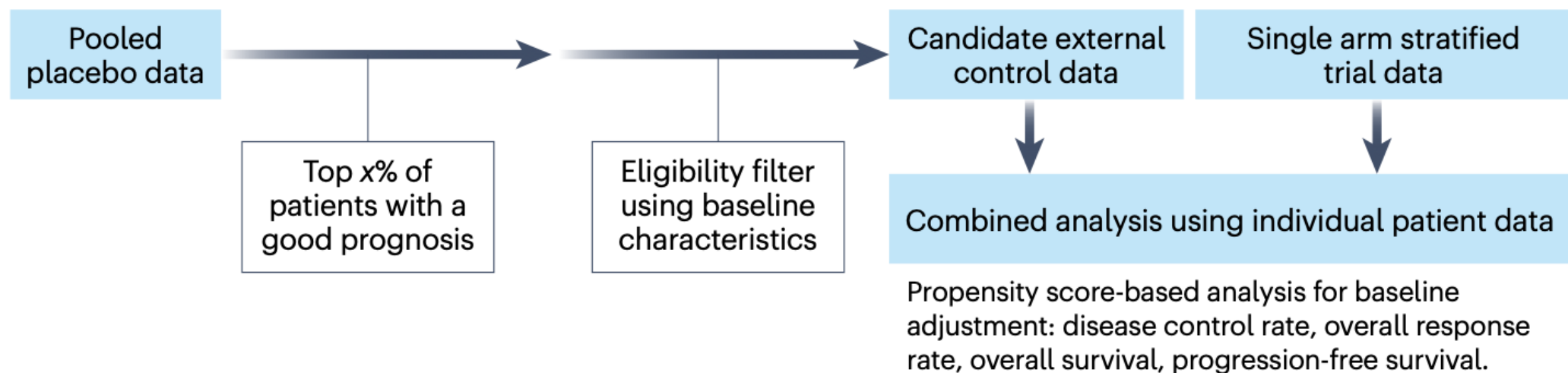
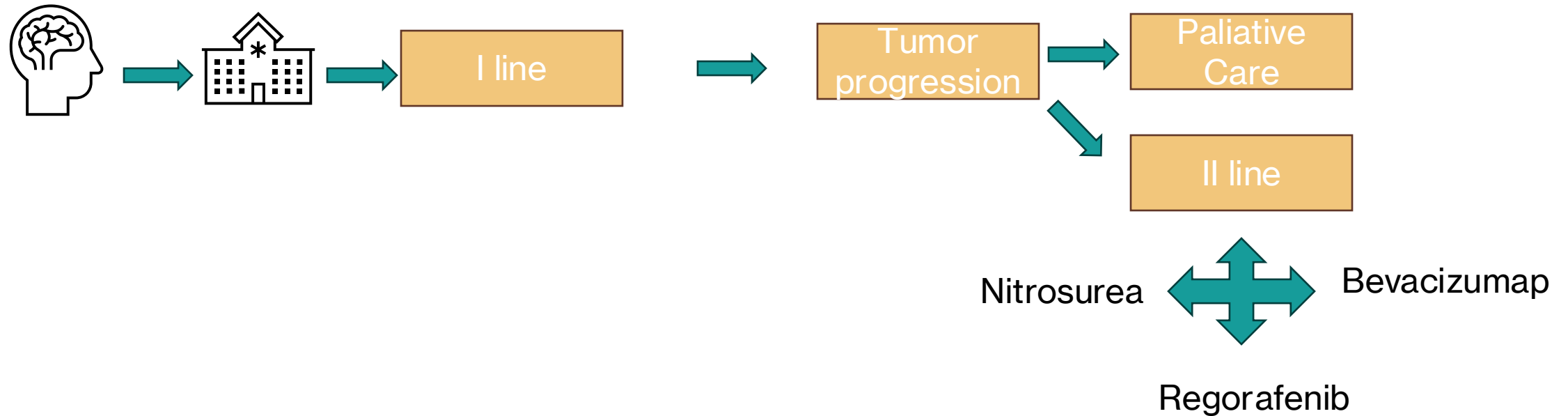


Fig. 1 | Three-step analysis for no-placebo initiative. First, participants enrolled in trials with placebo arms will be selected based on compatible patient demographics and key characteristics. These data will form the synthetic control arm. Second,

patients in the top percentile for overall survival will be extracted from the synthetic control arm. Third, the synthetic control arm will be compared with patients in the trial, using propensity scored-based analysis.

CAN MOLECULAR STATUS GUIDE OPTIMAL TREATMENT SELECTION IN RECURRENT GLIOBLASTOMA?

Generate Synthetic Arm



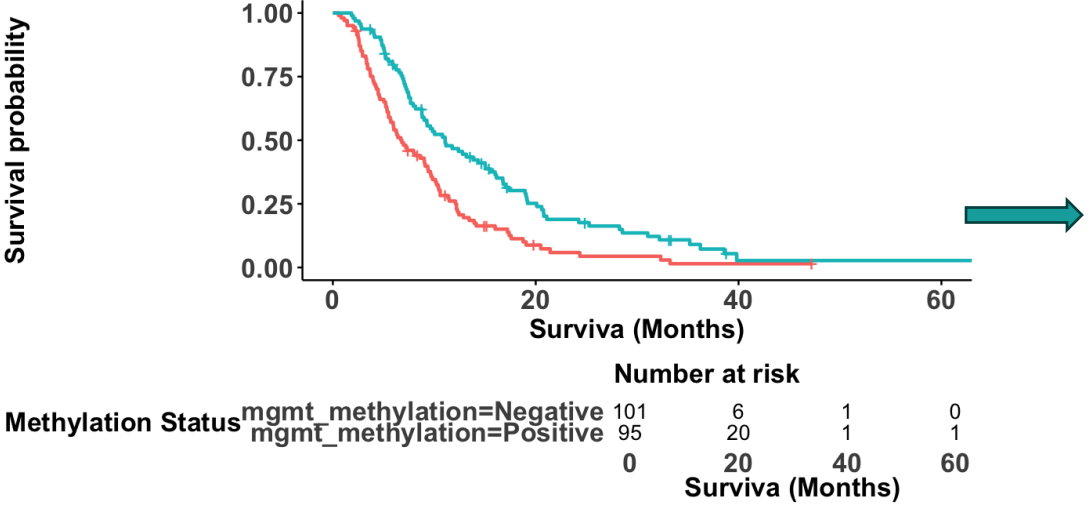
Dr. Lombardi G.

ORIGINAL VS SYNTHETIC: DO THE DATA TELL THE SAME STORY?

STATISTIC PROPRIETY

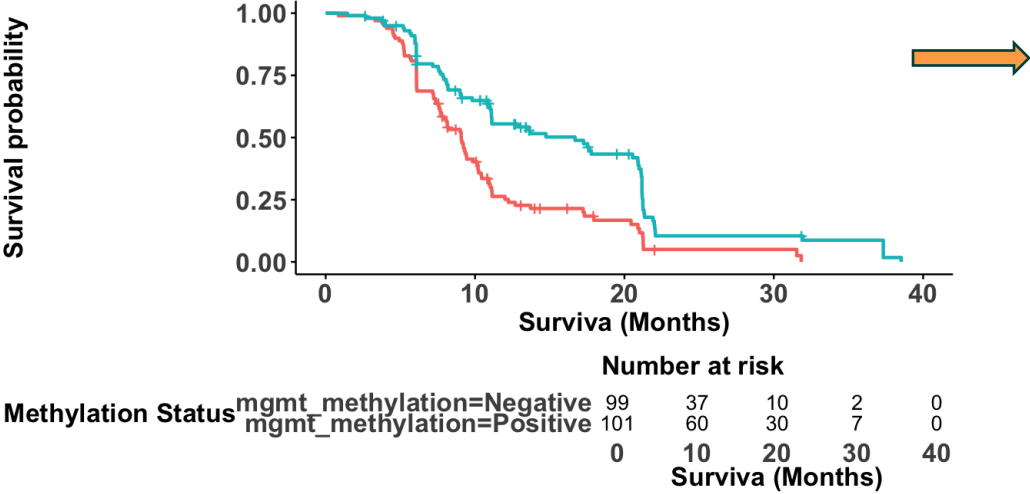
Survival by mgmt in original Data

Methylation Status + mgmt_methylation=Negative - mgmt_methyl

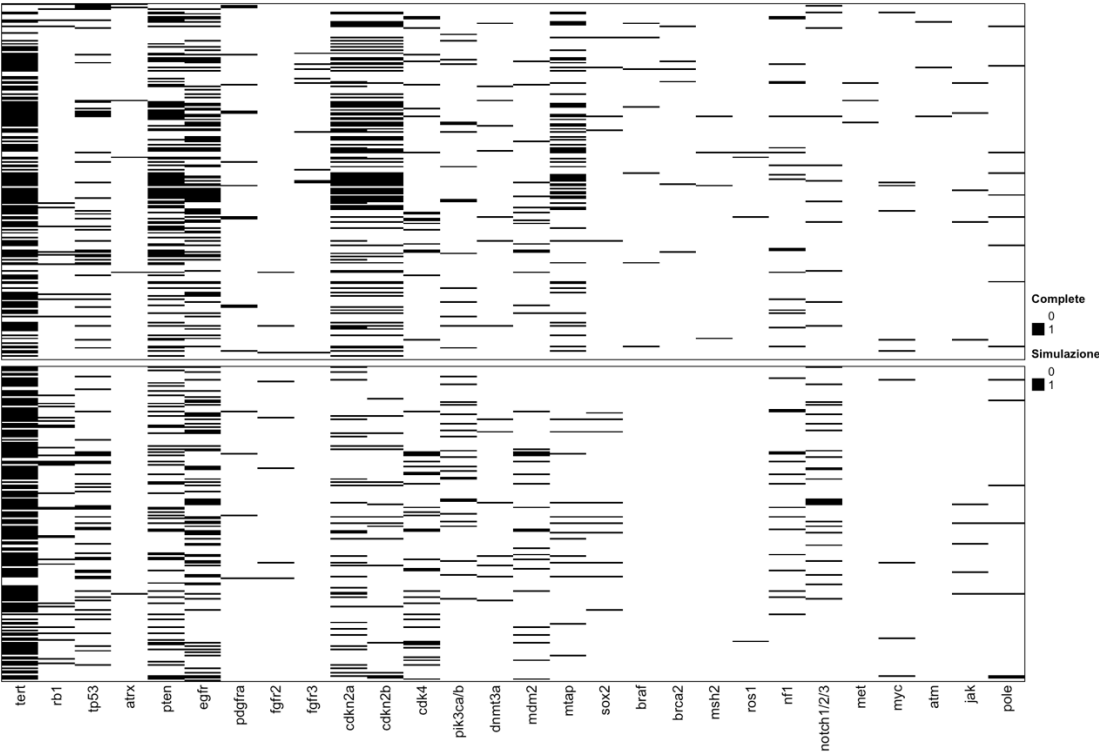


Survival by mgmt in Synthetic Data

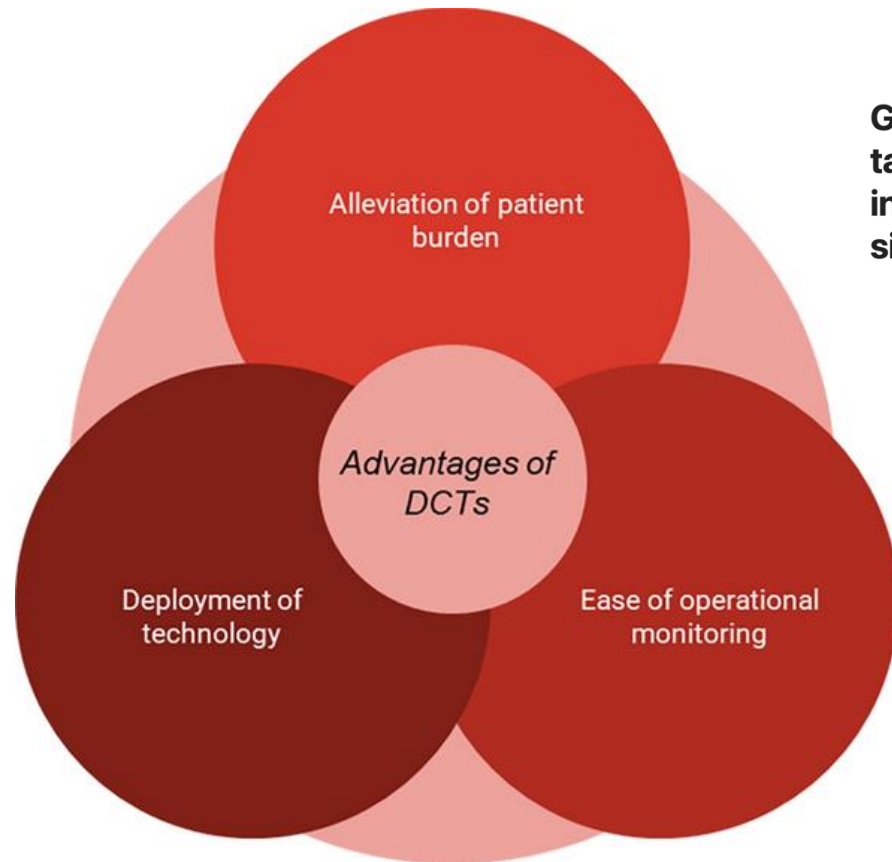
Methylation Status + mgmt_methylation=Negative - mgmt_methyl



GENOMIC LANDASCAPE



METHODOLOGICAL INNOVATIONS IN AI-DRIVEN TRIALS: SYNTHETIC ARMS, CAUSAL ESTIMATION, AND DCT FRAMEWORKS



G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study

Together, these methods lay the foundation for a new generation of clinical trials – flexible, inclusive, and analytically robust – where AI, through techniques like Double Machine Learning and model averaging, becomes a powerful engine for causal inference in both real-world and synthetic data.

SPRINGER NATURE Link

Find a journal Publish with us Track your research Search

Home > Therapeutic Innovation & Regulatory Science > Article

The Next Horizon of Drug Development: External Control Arms and Innovative Tools to Enrich Clinical Trial Data

Review | Open access | Published: 25 March 2024

Volume 58, pages 443–455, (2024) [Cite this article](#)

Model Averaging and Double Machine Learning

Achim Ahrens, Christian B. Hansen, Mark E. Schaffer, Thomas Wiemann

This paper discusses pairing double/debiased machine learning (DDML) with stacking, a model averaging method for combining multiple candidate learners, to estimate structural parameters. In addition to conventional stacking, we consider two stacking variants available for DDML: short-stacking exploits the cross-fitting step of DDML to substantially reduce the computational burden and pooled stacking enforces common stacking weights over cross-fitting folds. Using calibrated simulation studies and two applications estimating gender gaps in citations and wages, we show that DDML with stacking is more robust to partially unknown functional forms than common alternative approaches based on single pre-selected learners. We provide Stata and R software implementing our proposals.

VIRTUAL MTB-DRIVEN HYPOTHESIS GENERATION

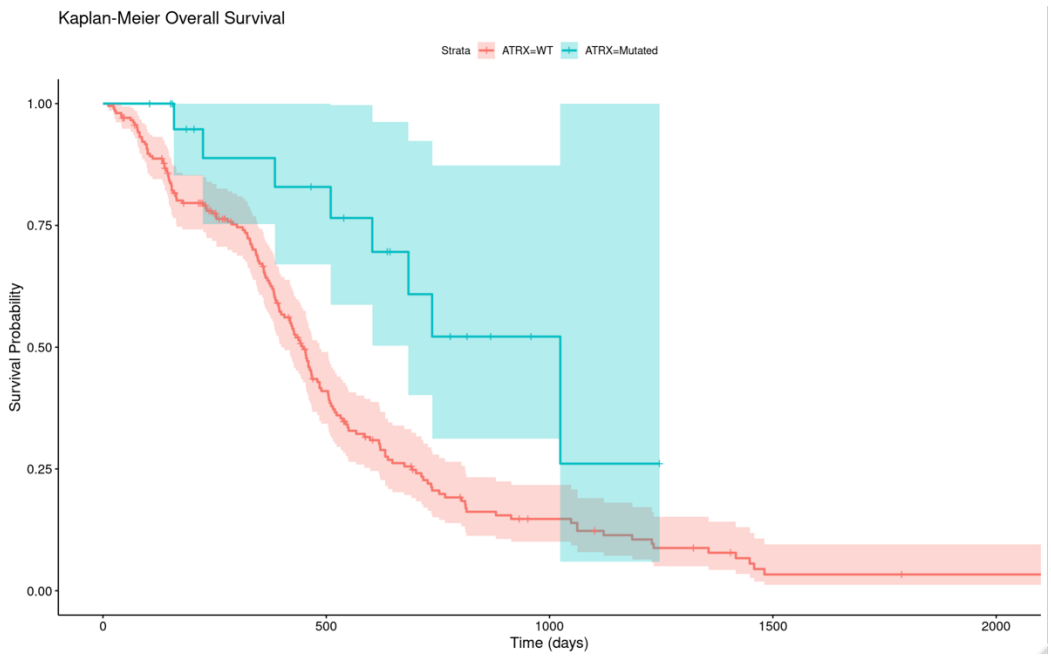
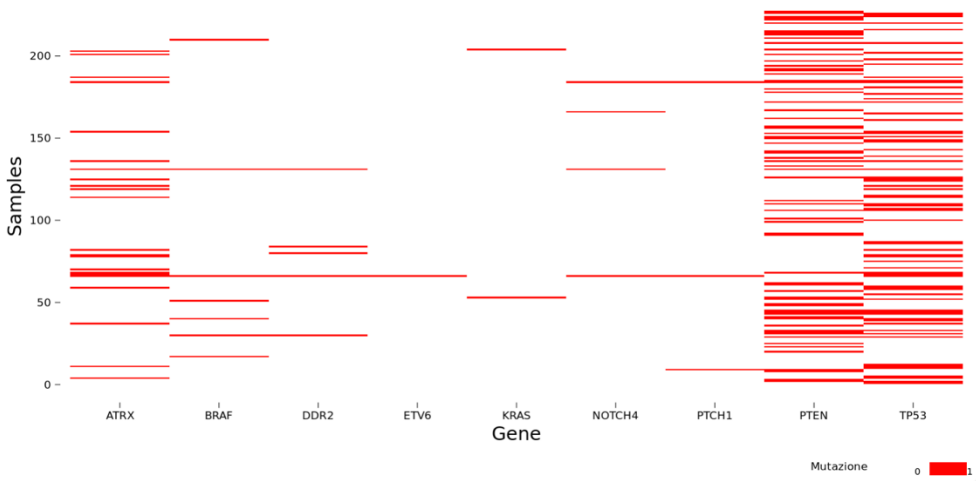
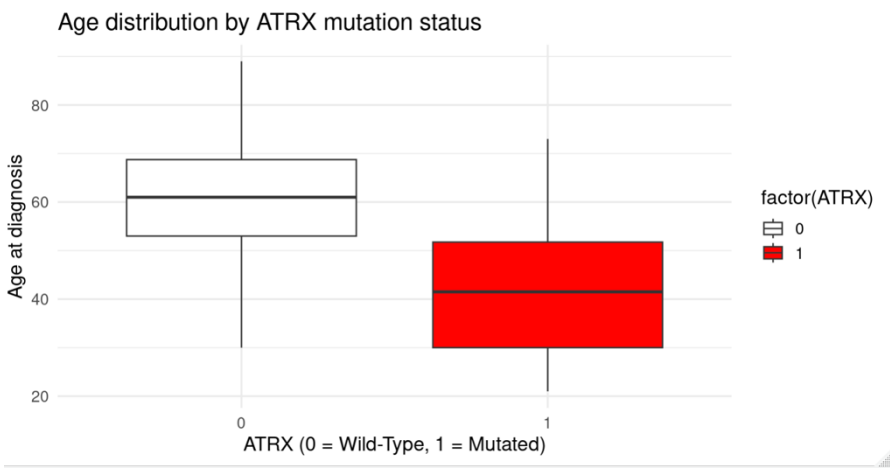
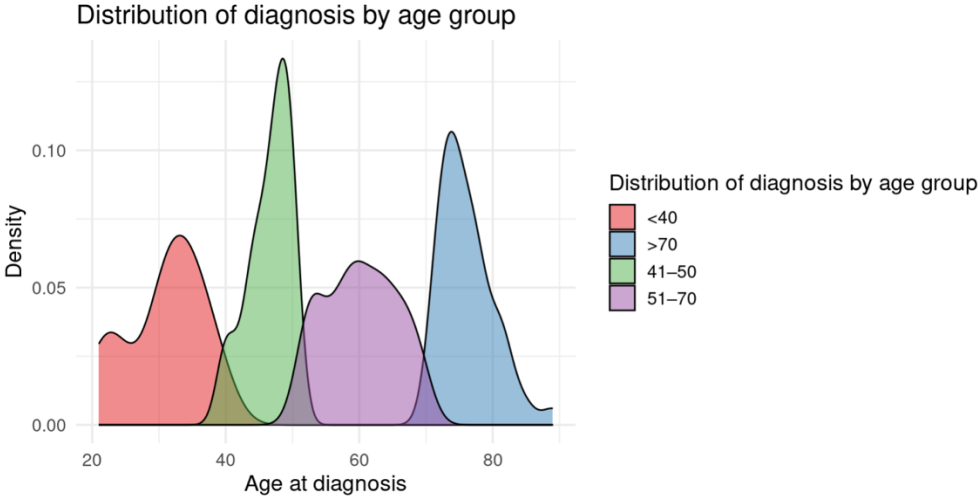
SYNTHETIC DATA GENERATION

Data Augmentation: Synthetic data can increase the volume and diversity of available datasets, aiding in the development of more robust AI models.

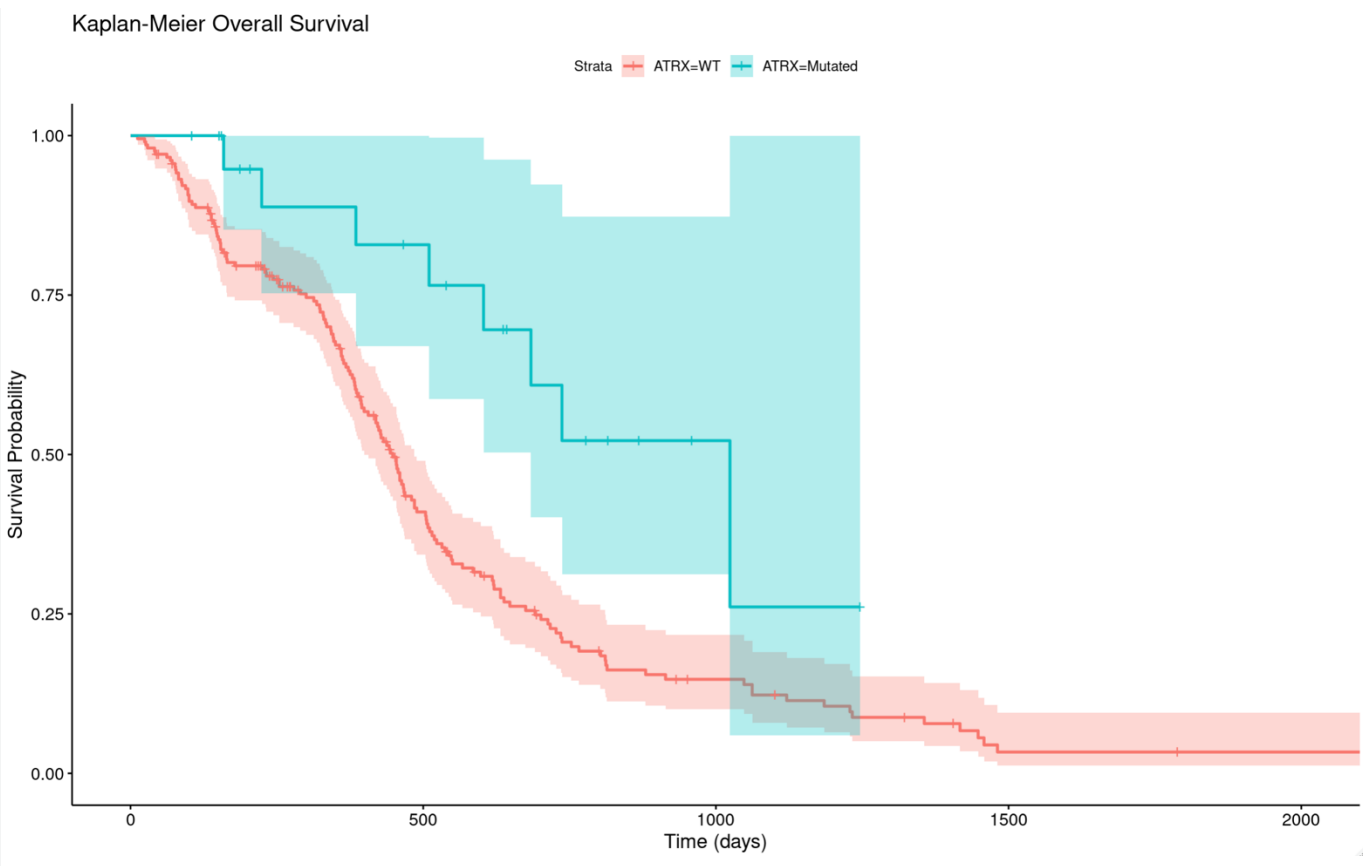
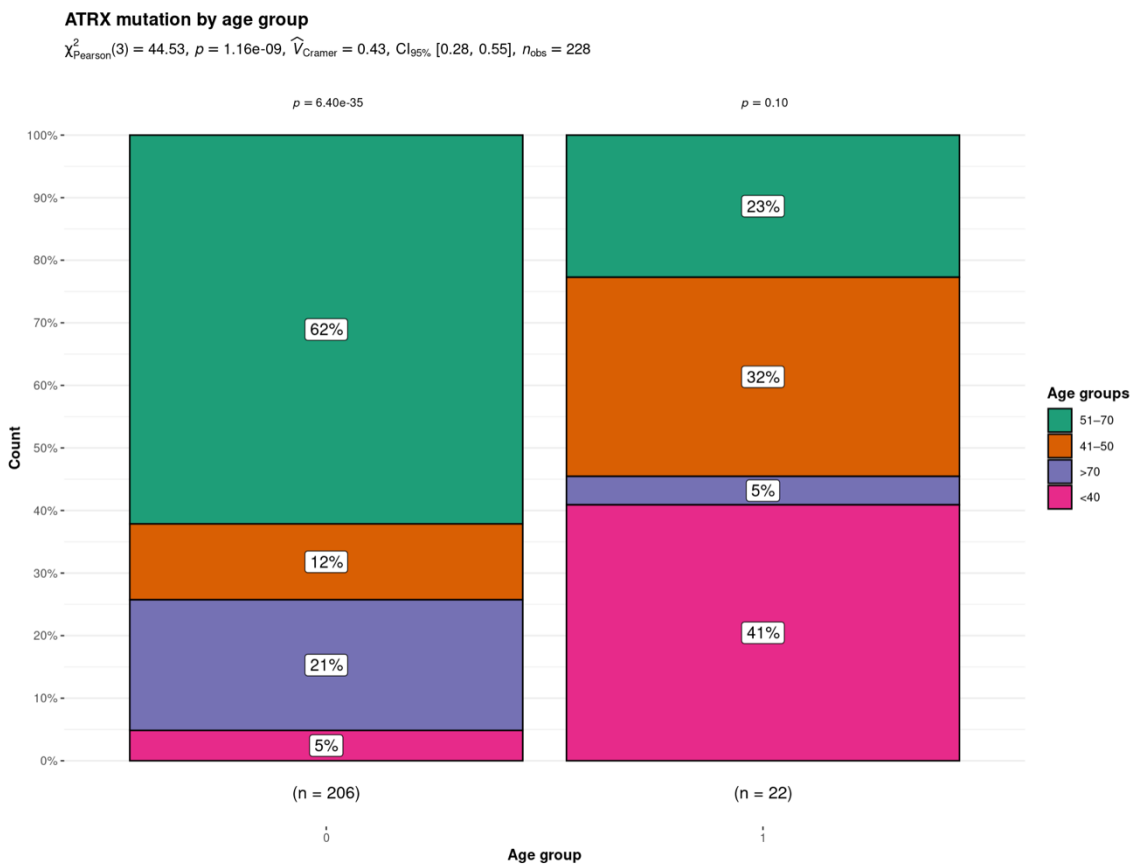
Privacy-Preserving Research: Synthetic data allows researchers to work with data that represents patient characteristics without revealing personal information, thus safeguarding patient privacy.

Standardization: Ensures consistent data formatting and structure, which is especially useful when integrating multiple data sources (e.g., genomic, clinical, radiomic data).

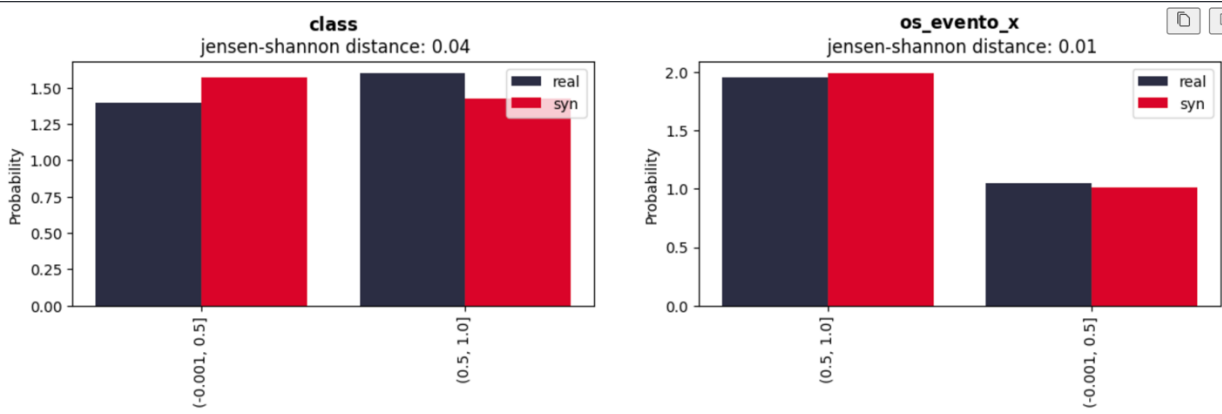
EXAMPLE OF AUGUMENTATION



EXAMPLE OF AUGUMENTATION

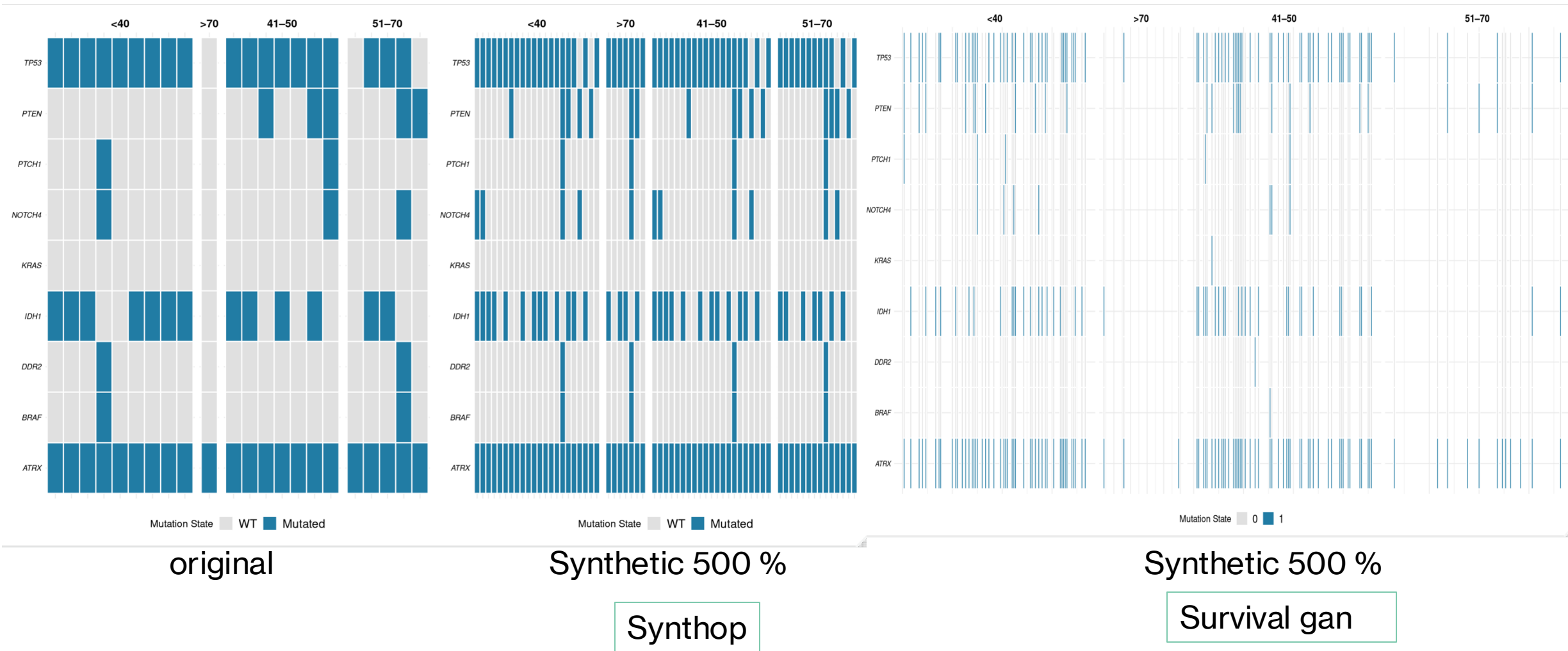


Comparatives			
	marginal_distributions	dummy_sampler	
sanity.data_mismatch.score	0.0 +/- 0.0	0.0 +/- 0.0	
sanity.common_rows_proportion.score	0.0 +/- 0.0	0.0 +/- 0.0	
sanity.nearest_syn_neighbor_distance.mean	0.562 +/- 0.0	0.415 +/- 0.093	
sanity.close_values_probability.score	0.125 +/- 0.0	0.438 +/- 0.062	
sanity.distant_values_probability.score	0.125 +/- 0.0	0.188 +/- 0.062	
stats.jensenshannon_dist.marginal	0.107 +/- 0.0	0.046 +/- 0.006	
stats.chi_squared_test.marginal	0.695 +/- 0.0	0.544 +/- 0.043	
stats.inv_kl_divergence.marginal	0.791 +/- 0.0	0.714 +/- 0.029	
stats.ks_test.marginal	0.487 +/- 0.0	0.795 +/- 0.028	
stats.max_mean_discrepancy.joint	0.25 +/- 0.0	0.266 +/- 0.016	
stats.wasserstein_dist.joint	43.831 +/- 0.0	13.554 +/- 0.427	
stats.prdc.precision	0.5 +/- 0.0	0.625 +/- 0.125	
stats.prdc.recall	1.0 +/- 0.0	0.938 +/- 0.062	
stats.prdc.density	0.35 +/- 0.0	0.438 +/- 0.062	
stats.prdc.coverage	0.75 +/- 0.0	0.938 +/- 0.062	
detection.detection_xgb.mean	0.833 +/- 0.0	0.602 +/- 0.093	
detection.detection_mlp.mean	0.611 +/- 0.0	0.583 +/- 0.028	
detection.detection_gmm.mean	0.611 +/- 0.0	0.444 +/- 0.028	
detection.detection_linear.mean	0.944 +/- 0.0	0.87 +/- 0.13	
privacy.k-anonymization.gt	999.0 +/- 0.0	999.0 +/- 0.0	
privacy.k-anonymization.syn	999.0 +/- 0.0	999.0 +/- 0.0	
privacy.k-map.score	0.0 +/- 0.0	0.0 +/- 0.0	
privacy.distinct l-diversity.gt	999.0 +/- 0.0	999.0 +/- 0.0	
privacy.distinct l-diversity.syn	999.0 +/- 0.0	999.0 +/- 0.0	
privacy.identifiability_score.score	0.25 +/- 0.0	0.125 +/- 0.0	
privacy.identifiability_score.score_OC	0.0 +/- 0.0	0.0 +/- 0.0	
privacy.DomiasMIA_BNAF.accuracy	0.205 +/- 0.0	0.667 +/- 0.026	
privacy.DomiasMIA_BNAF.aucroc	0.452 +/- 0.0	0.906 +/- 0.029	



	marginal_distributions	dummy_sampler
sanity.data_mismatch.score	0.0 +/- 0.0	0.0 +/- 0.0
sanity.common_rows_proportion.score	0.0 +/- 0.0	0.023 +/- 0.023
sanity.nearest_syn_neighbor_distance.mean	0.551 +/- 0.0	0.155 +/- 0.019
sanity.close_values_probability.score	0.136 +/- 0.0	0.841 +/- 0.023
sanity.distant_values_probability.score	0.182 +/- 0.0	0.068 +/- 0.023
stats.jensenshannon_dist.marginal	0.048 +/- 0.0	0.026 +/- 0.006
stats.chi_squared_test.marginal	0.843 +/- 0.0	0.72 +/- 0.032
stats.inv_kl_divergence.marginal	0.916 +/- 0.0	0.881 +/- 0.001
stats.ks_test.marginal	0.758 +/- 0.0	0.881 +/- 0.023
stats.max_mean_discrepancy.joint	0.123 +/- 0.0	0.122 +/- 0.001
stats.wasserstein_dist.joint	1.539 +/- 0.0	0.547 +/- 0.1
stats.prdc.precision	1.0 +/- 0.0	1.0 +/- 0.0
stats.prdc.recall	1.0 +/- 0.0	1.0 +/- 0.0
stats.prdc.density	0.718 +/- 0.0	0.945 +/- 0.036
stats.prdc.coverage	0.818 +/- 0.0	1.0 +/- 0.0
detection.detection_xgb.mean	0.842 +/- 0.0	0.594 +/- 0.064
detection.detection_mlp.mean	0.676 +/- 0.0	0.603 +/- 0.026
detection.detection_gmm.mean	0.36 +/- 0.0	0.475 +/- 0.025
detection.detection_linear.mean	0.863 +/- 0.0	0.591 +/- 0.122
privacy.delta-presence.score	3.0 +/- 0.0	1.667 +/- 0.667
privacy.k-anonymization.gt	7.0 +/- 0.0	7.0 +/- 0.0
privacy.k-anonymization.syn	7.0 +/- 0.0	4.0 +/- 3.0
privacy.k-map.score	5.0 +/- 0.0	5.0 +/- 2.0
privacy.distinct l-diversity.gt	7.0 +/- 0.0	7.0 +/- 0.0
privacy.distinct l-diversity.syn	7.0 +/- 0.0	4.0 +/- 3.0
privacy.identifiability_score.score	0.318 +/- 0.0	0.409 +/- 0.045
privacy.identifiability_score.score_OC	0.045 +/- 0.0	0.364 +/- 0.091
privacy.DomiasMIA_BNAF.accuracy	0.542 +/- 0.0	0.536 +/- 0.005
privacy.DomiasMIA_BNAF.aucroc	0.634 +/- 0.0	0.63 +/- 0.001

SYNTHETIC MUTATION PROFILES: PRESERVING SIGNAL OR INTRODUCING NOISE?

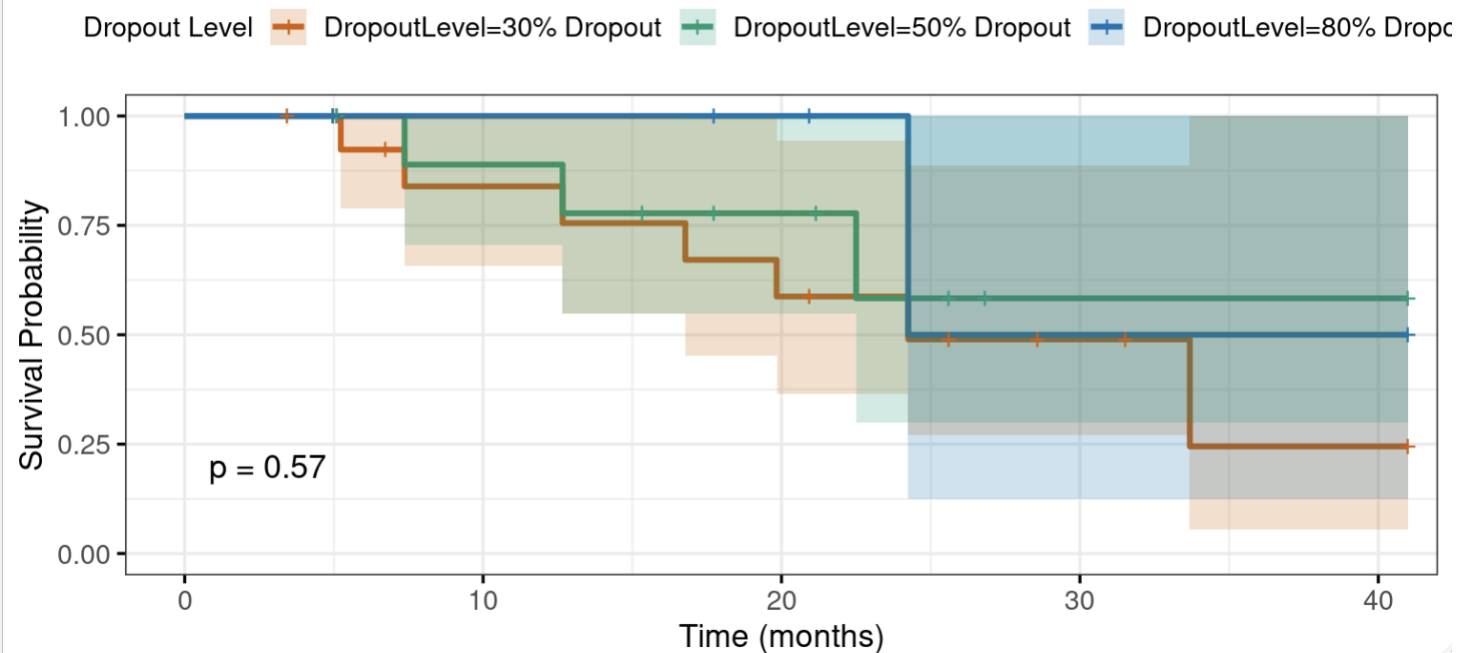


WHAT HAPPENS WHEN CRITICAL MOLECULAR INFORMATION IS LOST?

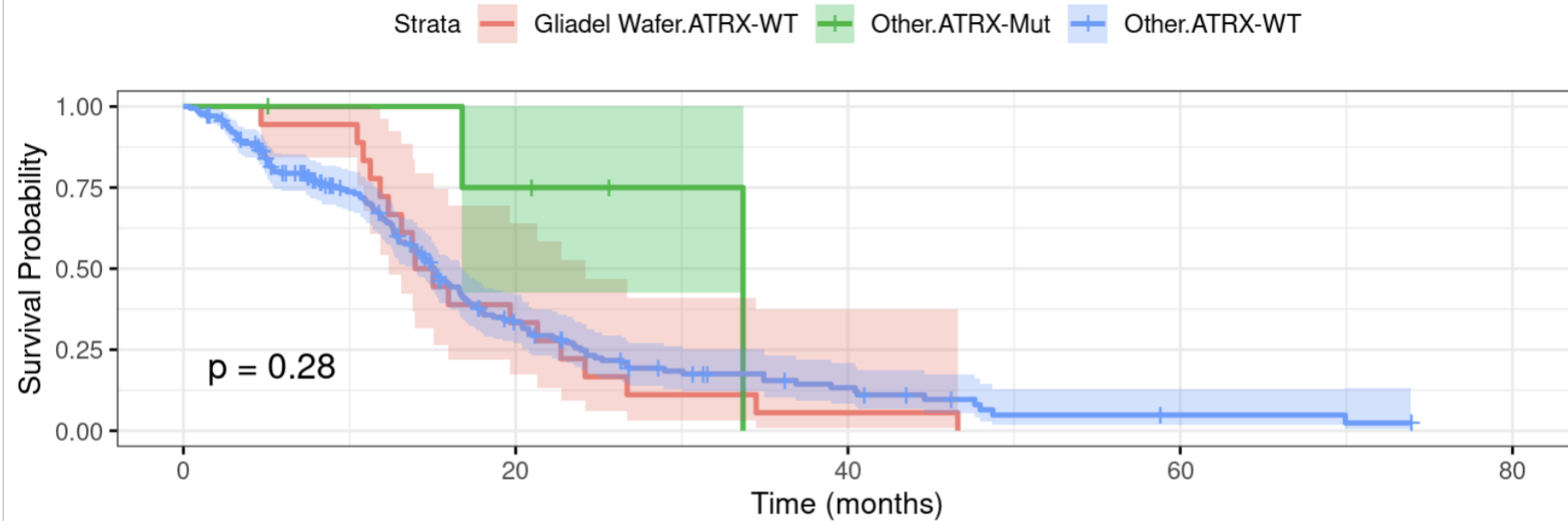
original

synthetic

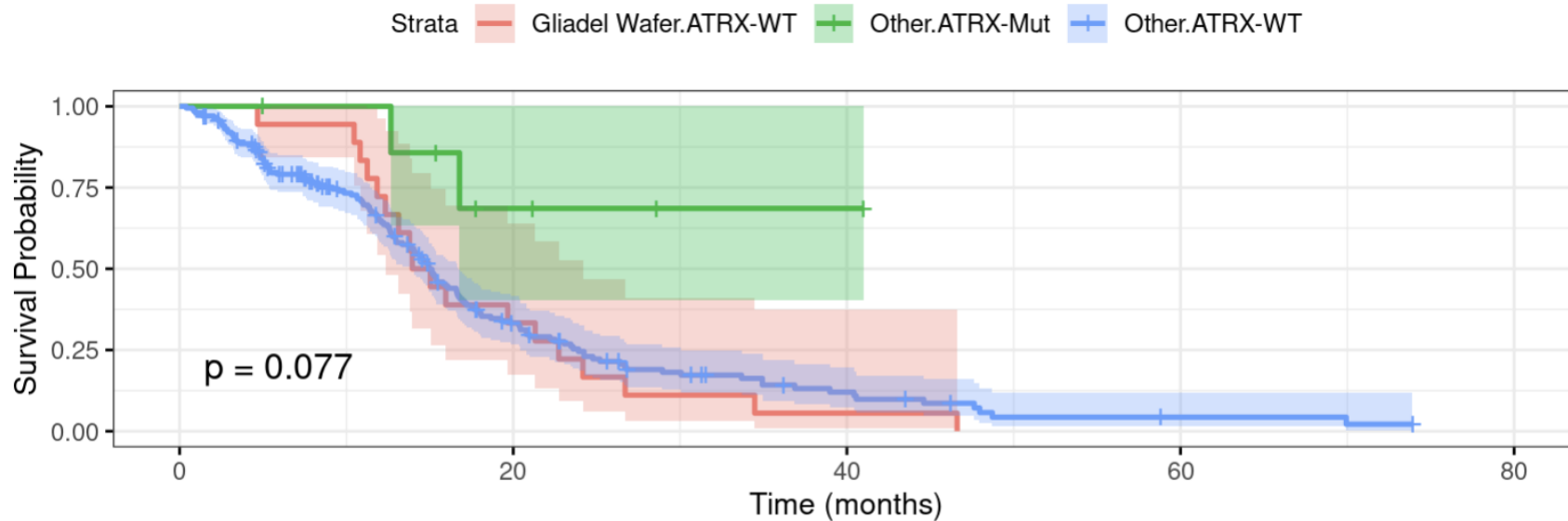
Kaplan-Meier: ATRX Mutated under Simulated Dropout



Kaplan-Meier: 70% dropout on ATRX



Kaplan-Meier: 50% dropout on ATRX



CONCLUSION

- **Synthetic data provides a safe sandbox** for scientific exploration, especially when real data is limited, sensitive, or restricted.
- **Context matters:** clinical relevance and underlying biological complexity must guide the generation and interpretation of synthetic datasets to avoid misleading conclusions.
- **Real data remains essential:** while synthetic data can support hypothesis generation, real-world data is the cornerstone for validation and clinical decision-making.

