Revolutionizing Healthcare with Synthetic Clinical Trial Data

Afrah Shafquat

Senior Data Scientist II, Medidata AI

May 10, 2023

MEDIDATA | ai

DASSAULT SYSTEMES | The 3DEXPERIENCE® Company

Synthetic Data Applications

Data sharing

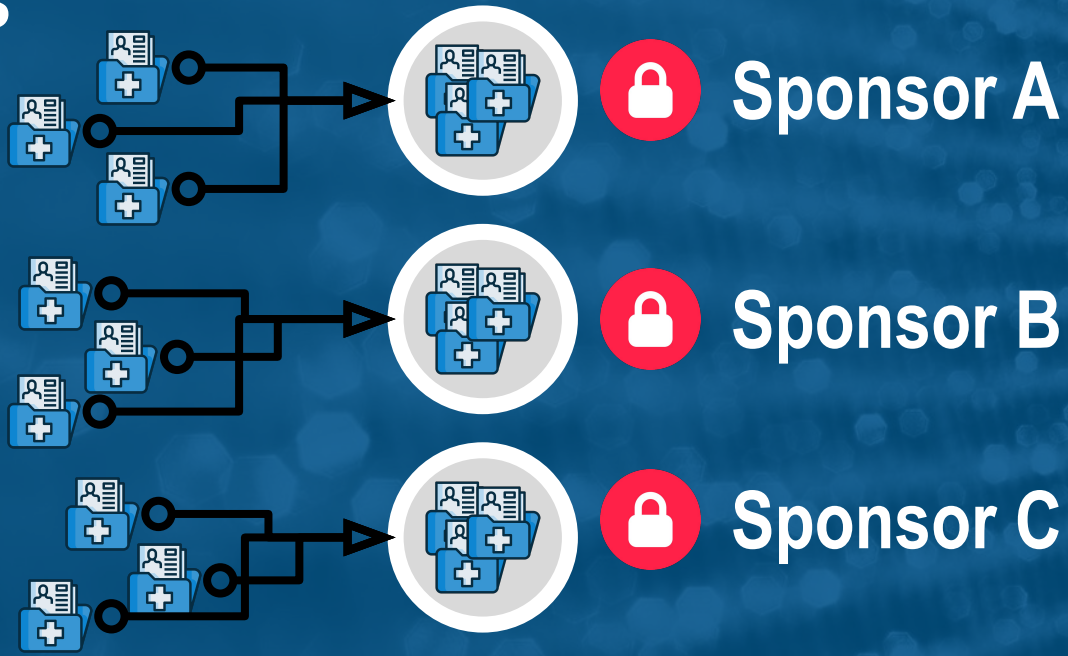
Data augmentation

Icons Source: Flaticon; Full credits in the last slide

Data Sharing

Icons Source: Flaticon; Full credits in the last slide

Clinical trial data remains siloed due to patient privacy and sponsor privacy concerns



MEDIDATA

Icons Source: Flaticon; Full credits in the last slide

DASSAULT SYSTEMES

Collaboration remains limited due to data access restrictions

Sponsor A

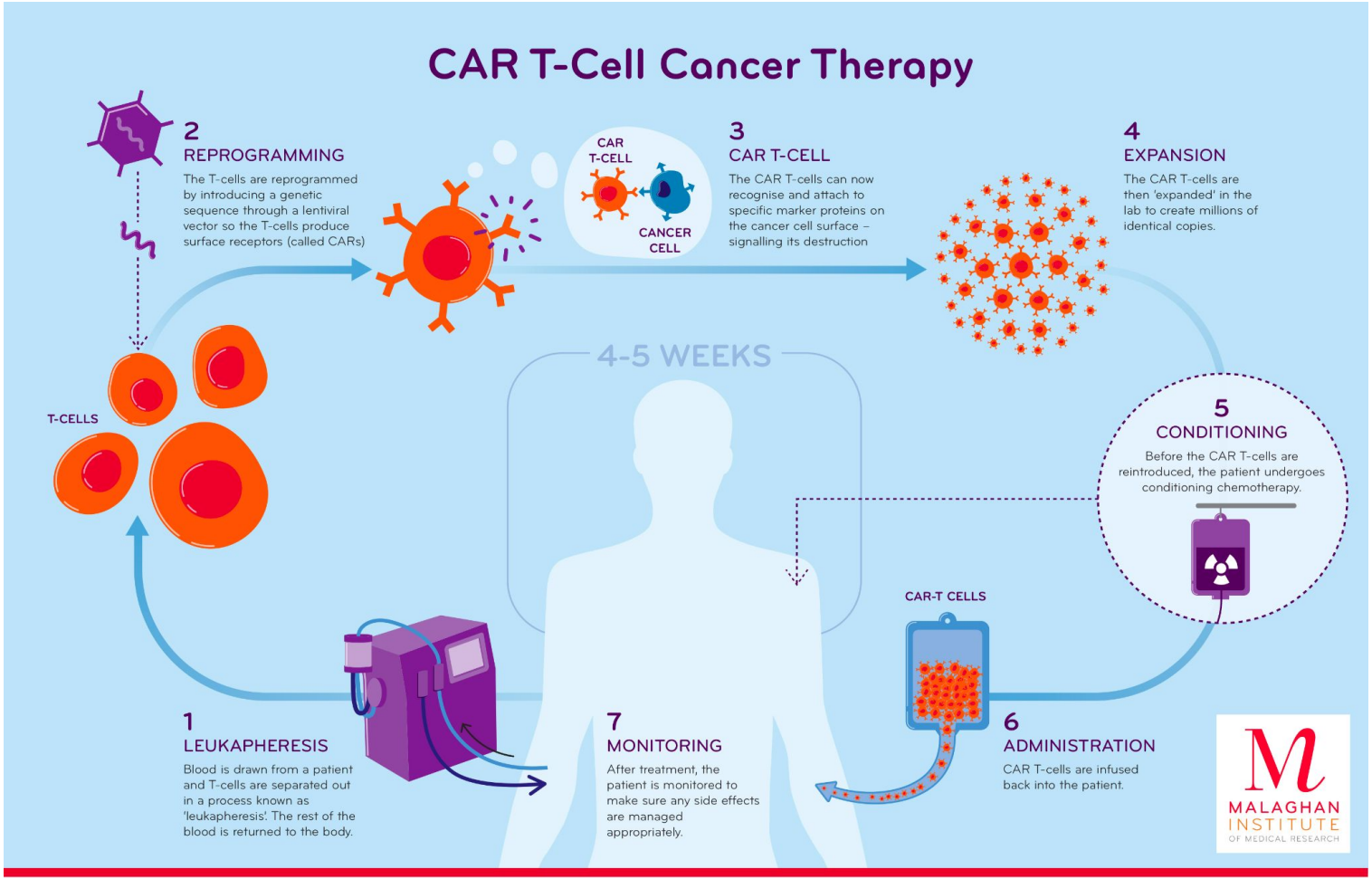
Sponsor B

Sponsor C

Icons Source: Flaticon; Full credits in the last slide

Why is collaboration necessary?

Icons Source: Flaticon; Full credits in the last slide



HEALTH • CANCER

Cancer's Newest Miracle Cure

August 21, 2017 issue of TIME.

MEDIDATA

DASSAULT SYSTEMES

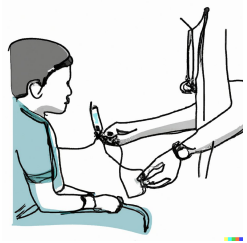
“Went looking for a miracle…”

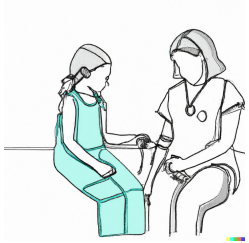
Source
CHOP News (May 2022)

MEDIDATA

DASSAULT SYSTEMES



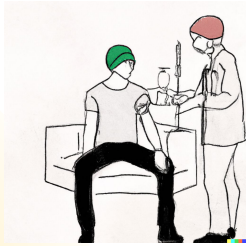


Cell therapy almost always begins its development journey by offering a treatment option to patients with very advanced & difficult to treat cancers.

* visuals created using DALL-E

No one can know exactly how any one patient will respond to these experimental treatments …

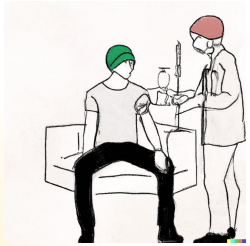
* visuals created using DALL-E



… but we do know that making a personalized medicine for an individual relies on our experience treating similar people with similar diseases

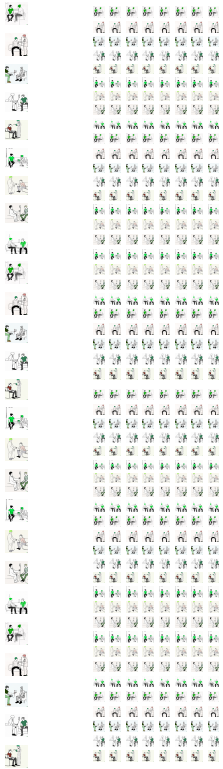
* visuals created using DALL-E

1 patient



1 doctor

1 trial



1 sponsor

Medidata AI

If we could learn from the combined experience of the whole world, rather than just relying on the experience of a single doctor or hospital or sponsor …

How well could we do?

IF WE™

* visuals created using DALL-E

Relapse

The journey of a CAR-T trial participant begins with the return of cancer …

* visuals created using DALL-E

Relapse

Enrollment

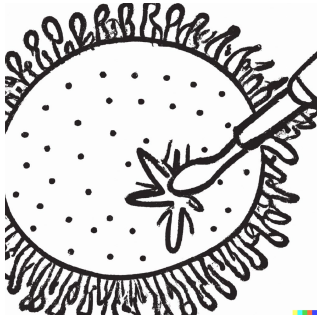
… the failed search for a viable approved treatment, and an introduction to a clinical investigator …

* visuals created using DALL-E

Relapse

Enrollment

Apheresis



Manufacturing

… removal of healthy lymphocytes and manufacture of a targeted dose of CAR-T cells …

* visuals created using DALL-E

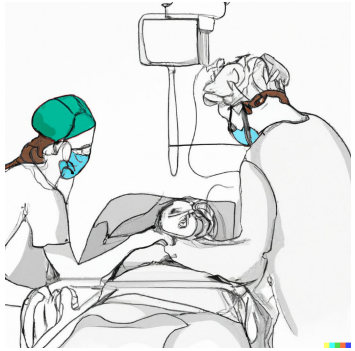
Relapse

Enrollment

Apheresis

Manufacturing

Treatment



Management

… treatment with investigational CAR-T therapy and careful management of patients through the side effects that follow …

* visuals created using DALL-E

Relapse

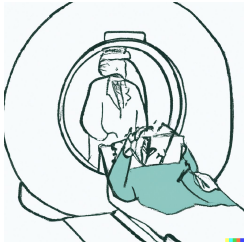
Enrollment

Apheresis

Manufacturing

Treatment

Management



Response

Control

… finally, a new response to treatment, disease control …

* visuals created using DALL-E

Relapse

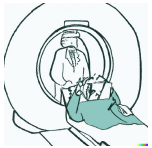
Enrollment

Apheresis

Manufacturing

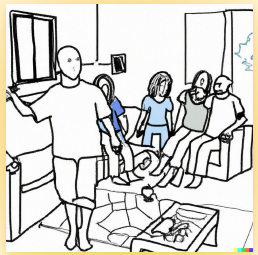
Treatment

Management



Response

Control



Home.

… and a return Home.

* visuals created using DALL-E

Relapse

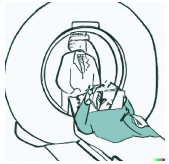
Enrollment

Apheresis

Manufacturing

Treatment

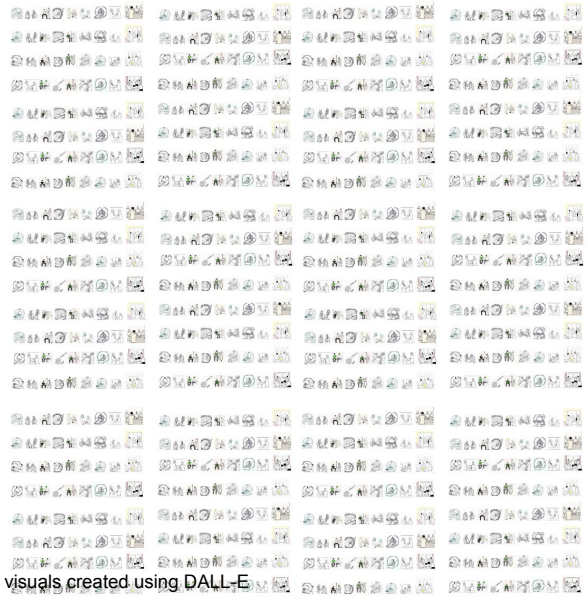
Management



Response

Control

Home.



Now, what if we understood this journey for 100 patients?

* visuals created using DALL-E

Relapse

Enrollment

Apheresis

Manufacturing

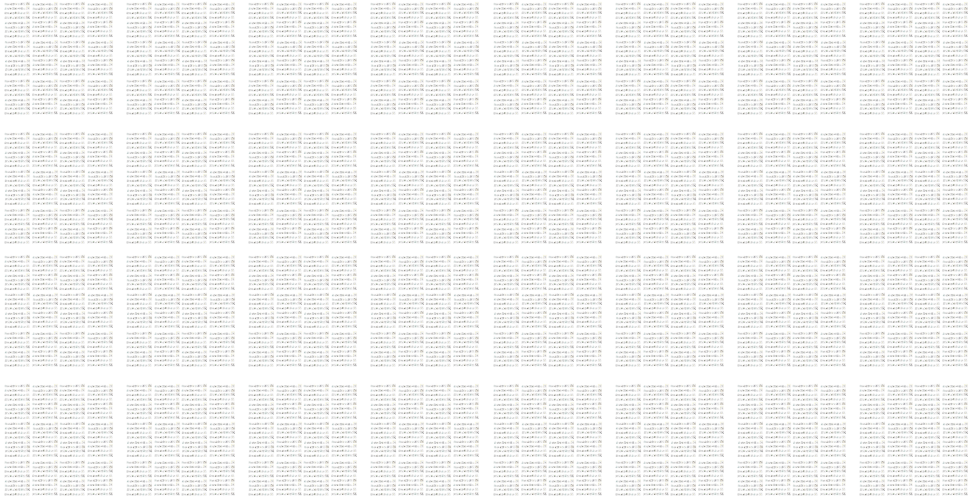
Treatment

Management

Response

Control

Home.



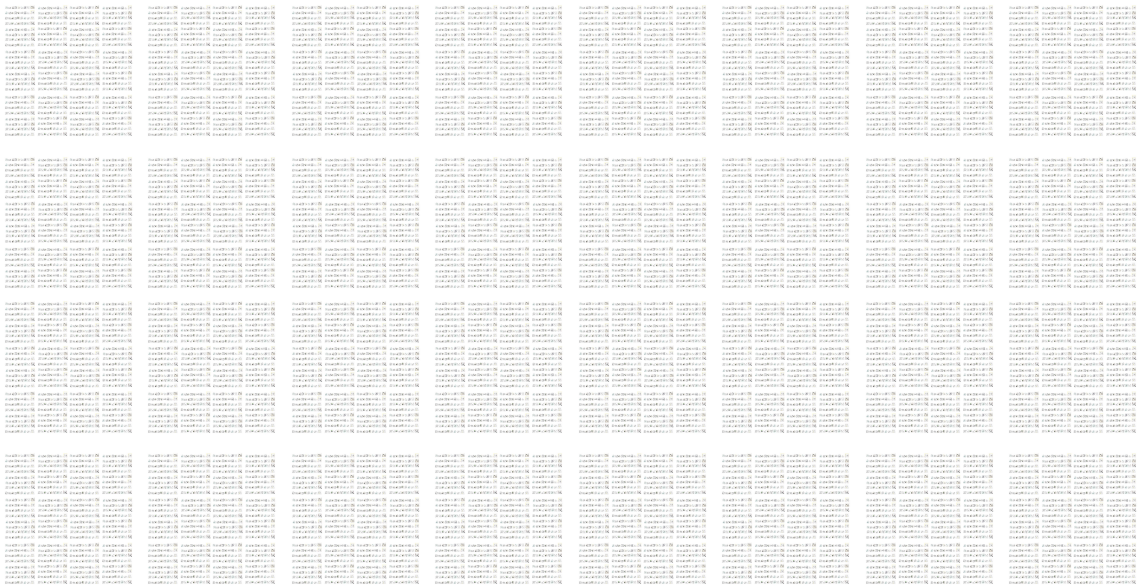
For 3,000?

* visuals created using DALL-E

1 patient. 1 journey. 1 experience.

Relapse | Enrollment | Apheresis | Manufacturing | Treatment | Management | Response | Control | Home.

Medidata AI — 3,000 CAR-T patient journeys today … and growing



* visuals created using DALL-E

If we could learn from the combined experience of the whole world, rather than just relying on the experience of a single doctor or hospital or sponsor, what?

How well could we do?

IF WE™

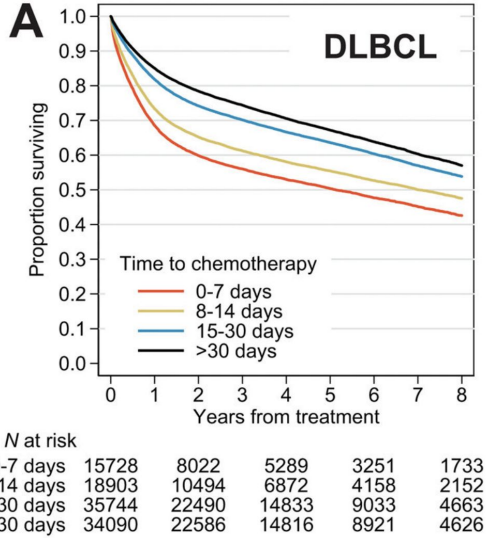
Relapse | Enrollment | Apheresis | Manufacturing | Treatment | Management | Response | Control | Follow-up

Every step in the patient's journey provides opportunities to design better trials to prove the benefit of new treatments

* visuals created using DALL-E

Relapse

Example



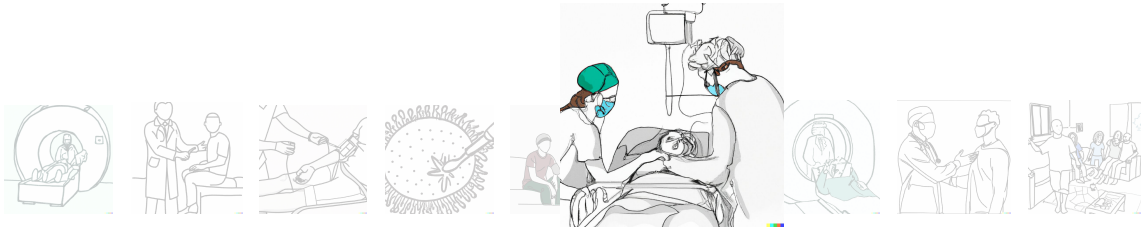
N at risk

0-7 days	15728	8022	5289	3251	1733
8-14 days	18903	10494	6872	4158	2152
15-30 days	35744	22490	14833	9033	4663
>30 days	34090	22586	14816	8921	4626

The length of **time between diagnosis and treatment** may be the **single strongest predictor of the prognosis** for trial participants with high risk lymphoma

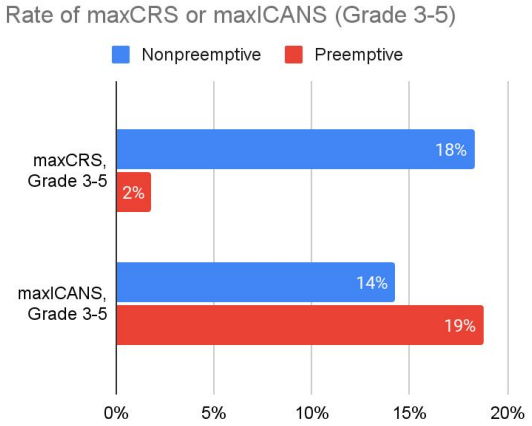
Alshaibani, A. et al (2019). *High risk patients with diffuse large B cell lymphoma are not enrolled on clinical trials.* JCO 37, e19058–e19058.

* visuals created using DALL-E



Management

Example



Pre-emptive treatment of CAR-T recipients with tocilizumab & dexamethasone reduces severe CRS rates but does not decrease rates of ICANS (neurotoxicity)

EHA 2023

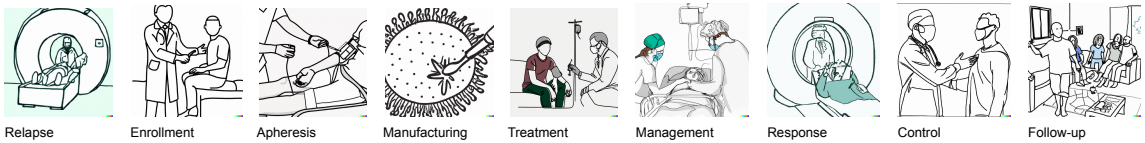
Authors:

Esther Nie, MD, PhD[2†]; Penelope Lafeuille, MS[1†]; Sheila Diamond, MS, CGC[1]; Jacob Aptekar, MD, PhD[1]; Vibhu Agarwal, PhD, MBA[1]

Affiliations:

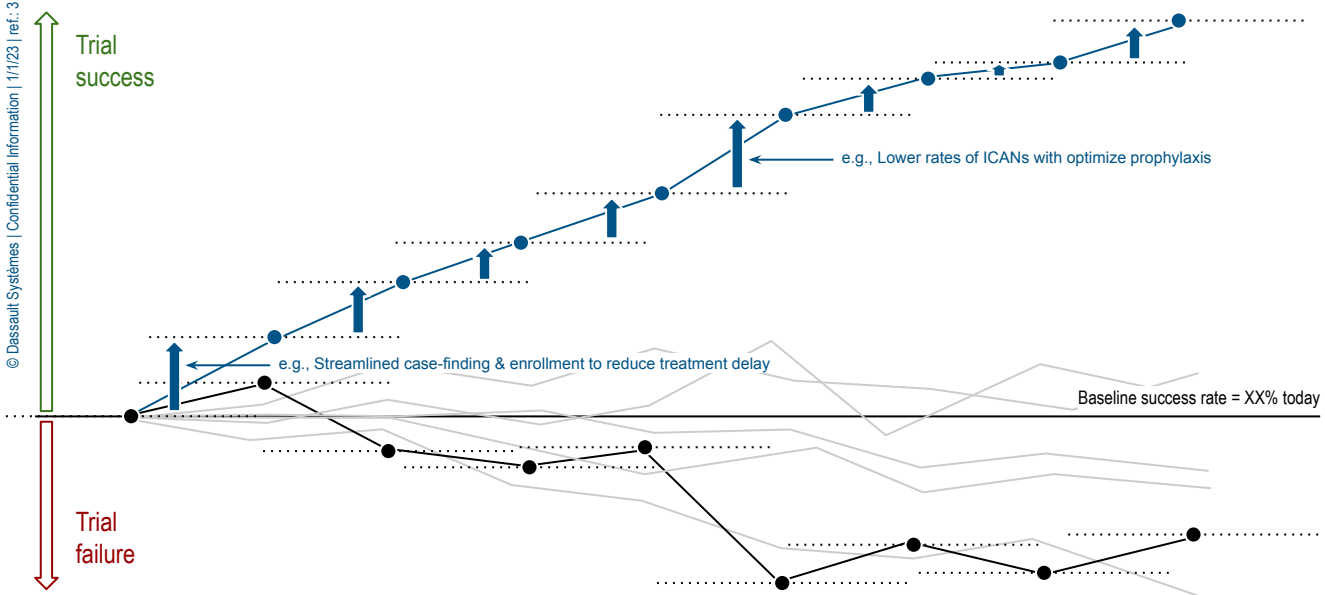
1. *Medidata, a Dassault Systèmes company, New York, NY*
2. *Stanford University, Stanford, CA*

* visuals created using DALL-E



Probability of Technical & Regulatory Success

Schematic



* visuals created using DALL-E

If we designed trials so that we made every decision informed by the experience of treating thousands of similar patients

How well could we do?

IF WE™

Select case examples and scientific publications

CASES

Top 20 BioPharma #1

Safer trials, better outcomes

Identified patients at high risk of developing severe CRS based on pre-infusion characteristics. Recommended optimal dosing regimen to minimize CRS by 20-30%

Top 20 BioPharma #2

Broader trial population, greater unmet need

Recommended eligibility criteria for 1L DLBCL trial to expand label and accelerate enrolment. Reduced trial timeline by 6-12 months

EU Biotech

Data-driven trial design, optimization

Developed optimal treatment strategies to minimize CRS and ICANs. 30% fewer cases of grade 3+ AEs. Identified patient likely to respond to therapy based on prior treatments

Top 20 BioPharma #3

Safer treatments in the clinic=bigger market

Predicted which patients are at higher risk of developing CRS based on prior treatment history and patient characteristics. ~20% higher accuracy in identifying high patients

PUBLICATIONS

ASCO

Predictors of severe CRS in longitudinal CAR T-cell clinical trial data

Poster presentation, 2022

American Society of Hematology

Deriving Predictive Features of Severe CRS from Pre-Infusion Clinical Data in CAR T-Cell Therapies

Poster presentation, 2022

NCCN

Evaluating Early Risk of Cytokine Release Syndrome (CRS) Induced by CAR T-Cell Therapy Using Pooled Clinical Trial Data

Poster presentation, 2022

EHA

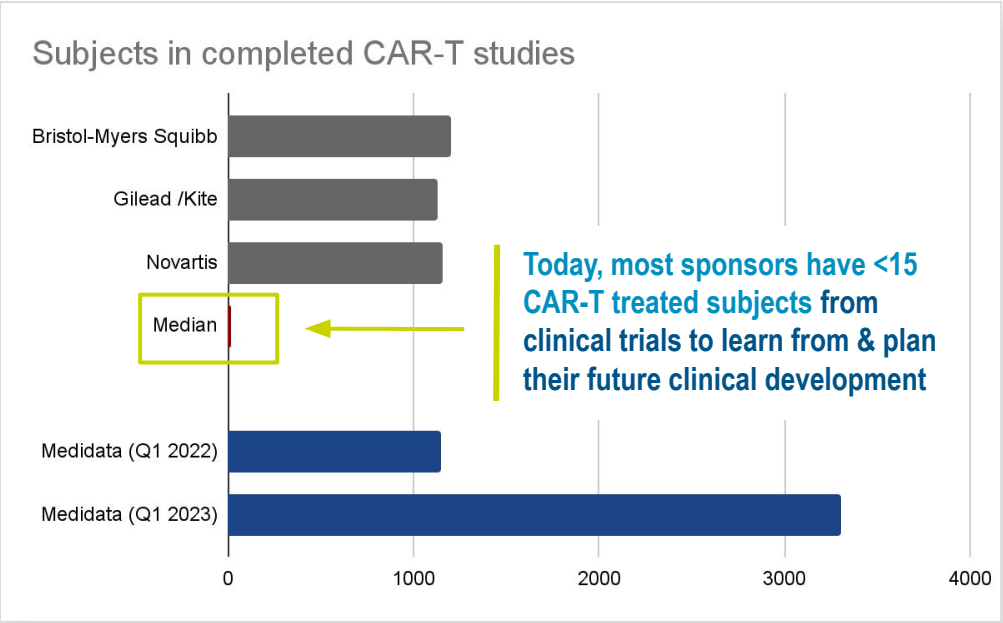
Cooccurrence patterns of CRS and ICANS in patients undergoing autologous CD19-targeted CAR T-cell treatments

In submission, 2023

MEDIDATA

DASSAULT SYSTEMES

60% of CAR-T data has been generated by 4 sponsors, and most sponsors working in this space have data on <15 subjects



50+ trials

CAR-T or T-Cell Engager (TCE) clinical trials run on Medidata's platform

>5k+ subjects

available to Medidata AI in unique CAR-T and TCE synthetic dataset

Synthetic data supports a healthy, resilient biopharma ecosystem

Forests rely on **nutrient sharing** by fungi between plants (*mycorrhiza*) to support a healthy, local ecosystem

Nutrients move between
plants via
a network of fungal hyphae (tubes) that digests, mobilizes & delivers
water & essential minerals to neighboring plants & trees
to enhance growth & disease resistance within
a small, geographical area

Mycorrhiza

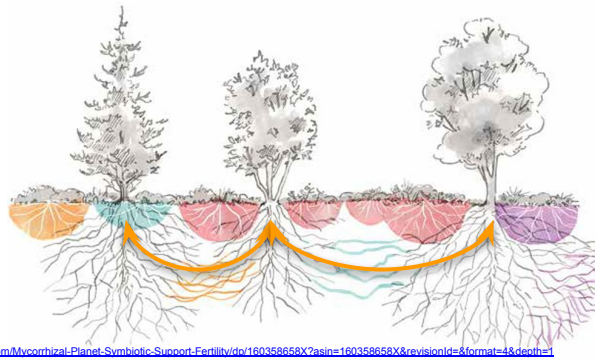
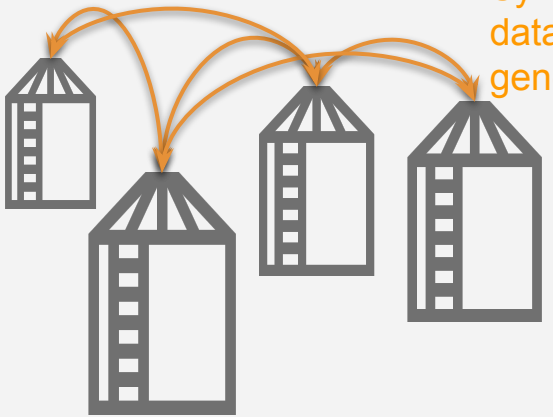


Image from book: https://www.amazon.com/Mycorrhizal-Planet-Symbiotic-Support-Fertility/dp/160358658X?asin=160358658X&revisionId=&format=4&depth=1

Biopharma relies on **knowledge sharing** to support a healthy, productive ecosystem for innovation

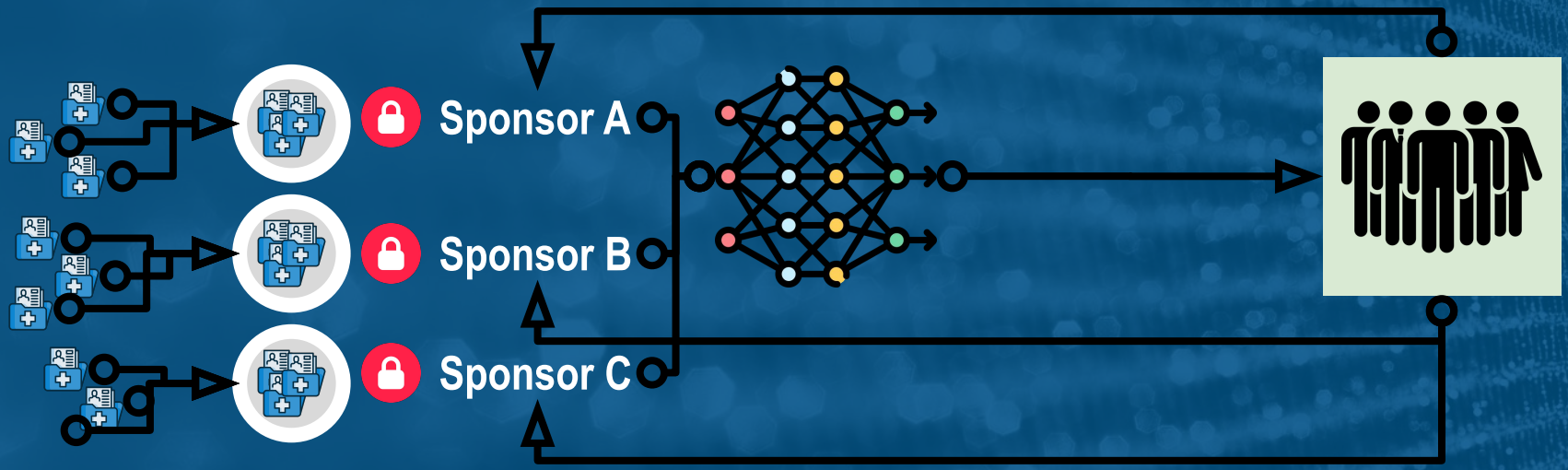
Insights move between
companies via
Synthetic data network that digests, mobilizes & delivers **information & insights** to PharmaCos & Biotechs
to develop safe, effective medicines within
a therapeutic area

Synthetic data generator



MEDIDATA

Synthetic data is a solution to encourage collaboration and continue innovation in pharma



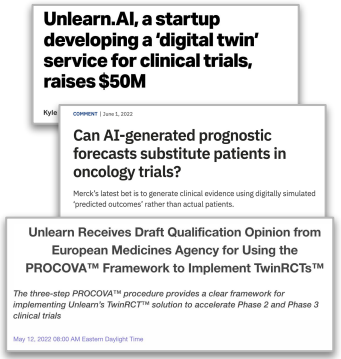
MEDIDATA

Icons Source: Flaticon; Full credits in the last slide

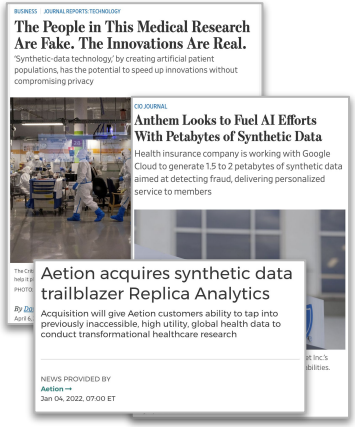
DASSAULT SYSTEMES

Synthetic Data landscape

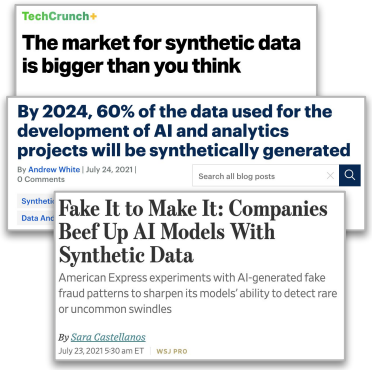
Clinical Trials



Real World Data



Industry at Large



MEDIDATA

DASSAULT SYSTEMES

Available open-source solutions

SmartNoise

MEDIDATA Simulants

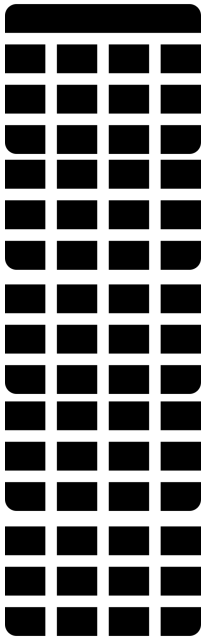
synthcity

gretel-synthetics by gretel™

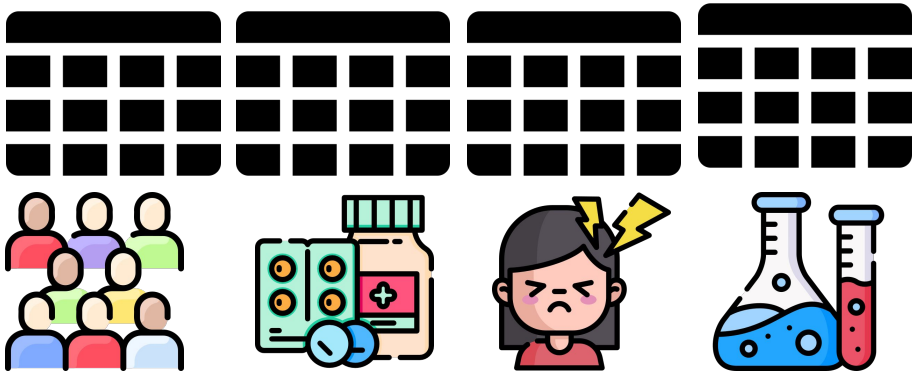
SDV
The Synthetic Data Vault

datacebo

Limitations in current solutions and the uniqueness of clinical trial data



vs.



Baseline | Medications | Adverse events | Labs

MEDIDATA

Icons Source: Flaticon; Full credits in the last slide

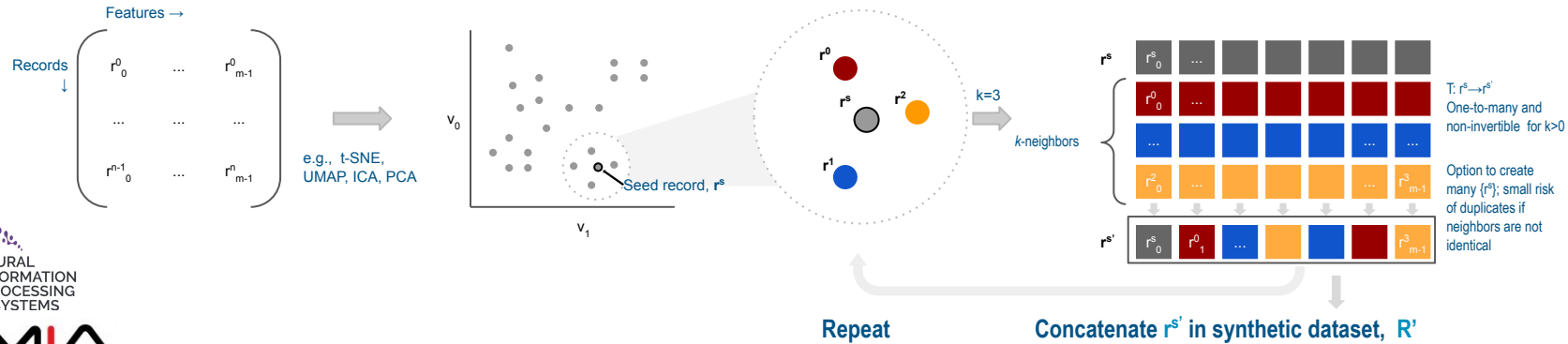
DASSAULT SYSTEMES

Simulants: Open-source Methodology

Privacy, fidelity & numerosity can be easily tuned and adjusted through choice of embedding, number of neighbors & feature linking

Simplified Process Diagram

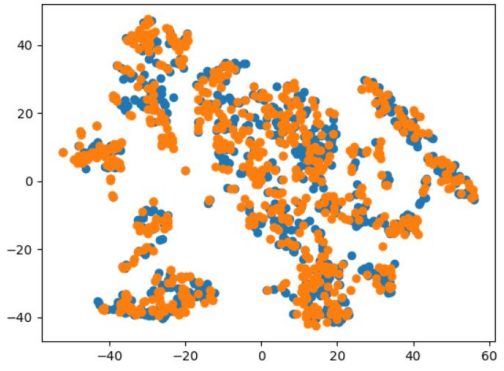
Load source dataset, R ➡ **Embed records, r^i, in metric space** ➡ **Select seed, r^s, and identify nearest neighbors** ➡ **Generate new record(s), $r^{s'}$, by randomly selecting features from *k*-neighbors**



1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis** - AMIA 2022
3. Open-source Simulants GitHUB: https://github.com/mdsol/Simulants
4. Simulants is implemented in Python and uses packages from open-source libraries including sciki-learn and numpy

Qualitative: t-SNE Embedding

t-SNE Embedding & Overlap



Alignment between source/template and Simulants generated synthetic dataset indicates synthesis reflects the source population well

Other open-source fidelity metrics available at the MIT/Data Cebo SDV project

SDV
The Synthetic Data Vault
datacebo

1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis -** AMIA 2022

MEDIDATA

Fidelity evaluation

DASSAULT SYSTEMES

Quantitative: Univariate measures

Comparison of Mean and other statistical similarity metrics

	Template	Simulants
NUM_CYC	7.030086	7.379656
TTE_SD1	56.193410	54.720630
TTE_DEATH	249.376791	254.849570
TTE_BOR	50.110315	49.786533
HEIGHT	161.403023	160.781633
TTE_PFS	179.895415	187.706304
TTE_GLYC_AE	85.014327	85.451289
PRIOR_RADIOTX_TIME	64.816619	68.999284
AGE	65.110315	64.769341
PRIOR_CANC_SURG_TIME	265.462751	230.846705
SLD_BASELINE	61.173352	60.402579
ECOG	0.638968	0.651862
TTE_PR1	62.312321	61.693410
DEPTH_RESP	-24.353725	-24.694126
TTE_SKIN_AE	24.965616	28.411175
PRIOR_CHEMO_TIME	659.383954	659.811605
TTE_CR1	52.563037	53.174785
TTE_PD1	179.385387	179.308023
TTE_2L	179.388252	182.510029

Comparison of mean across numerical features shows the mean of source/template and Simulants generated synthetic data are close.

Other open-source fidelity metrics available at the MIT/Data Cebo SDV project

SDV
The Synthetic Data Vault
datacebo

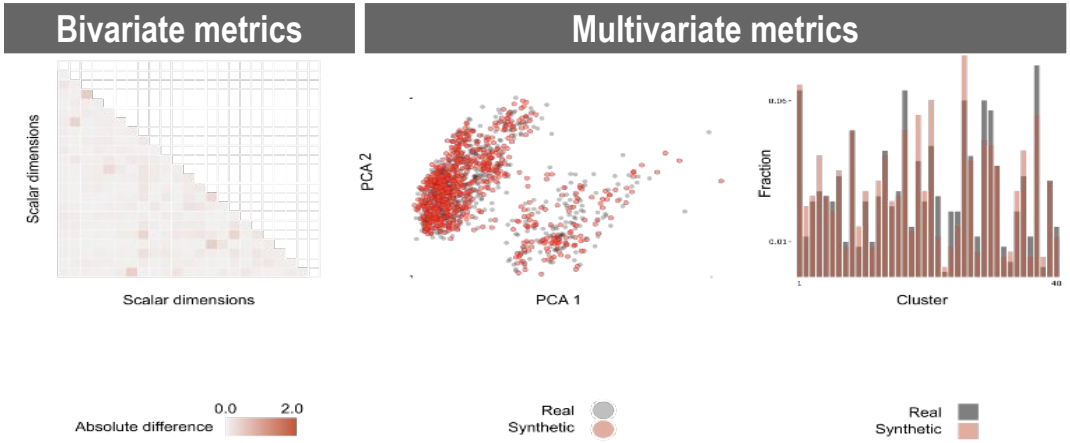
1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis -** AMIA 2022

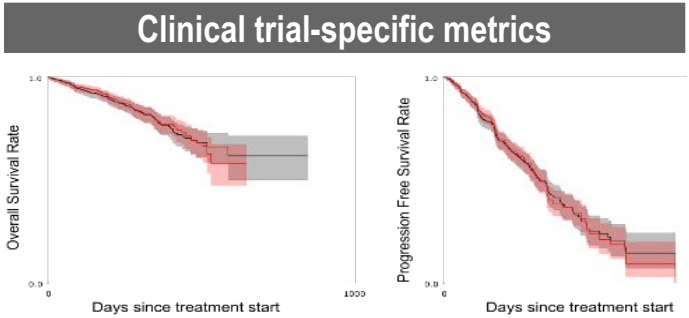
MEDIDATA

Fidelity evaluation

DASSAULT SYSTEMES

Quantitative metrics





Bivariate metrics like comparison of pairwise correlations allow assessment of preservation of correlations across features*

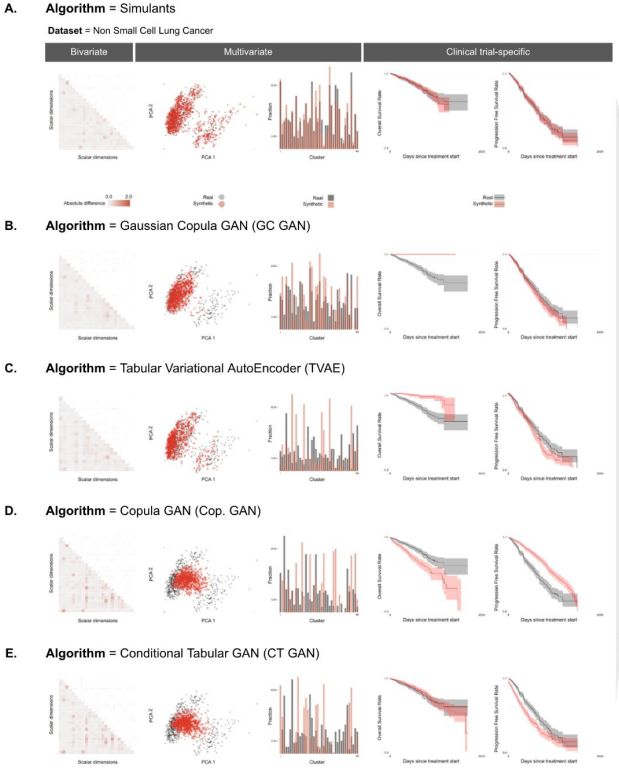
* All fidelity metrics were computed using open-source packages including scipy, tsne, scikit-learn, numpy etc.

Clinical trial specific metrics like comparison of survival probability as computed using Kaplan-Meier curves allow assessment of preservation of clinical insights in synthetic data

1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis -** AMIA 2022

Results – Fidelity & cross validation

Benchmarks – Simulants outperforms all methods in Clinical Trial data



F. Fidelity quantification summary

UPDATE

Dataset	Algorithm	Univariate: Dimensions where K-S, Chi-Square p<0.05	Bivariate: Correlation, mean absolute difference	Multivariate: Silhouette score	Multivariate: "Bag of Words" distance	Clinical trial - specific: log-rank Overall Survival K-M	Clinical trial - specific: log-rank Progression Free Survival K-M
	Metric						
	Best	0/N	0	0	0	1	1
	Worst	N/N	1	1		0	0
Non small cell lung cancer	Simulants	0/171	0.11	0.000	0.042	0.47	0.89
	GC GAN	56/171	0.08	0.004	0.106	8.75E-26	0.46
	TVAE	98/171	0.10	0.014	0.215	5.62E-16	0.39
	Cop. GAN	64/171	0.22	0.040	0.236	9.63E-08	9.03E-09
	CTGAN	58/171	0.20	0.027	0.242	0.56	0.6
Diffuse Large B Cell Lymphoma	Simulants	2/174	0.07	0.000	0.024	0.99	NA
	GC GAN	88/174	0.09	0.009	0.210	3.31E-35	NA
	TVAE	94/174	0.12	0.026	0.159	3.82E-09	NA
	Cop. GAN	82/174	0.22	0.014	0.330	2.49E-04	NA
	CTGAN	83/174	0.22	0.018	0.320	3.02E-11	NA
Acute Lymphoblastic Leukemia	Simulants	11/142	0.08	0.000	0.013	NA	NA
	GC GAN	72/142	0.20	0.032	0.261	NA	NA
	TVAE	73/142	0.48	0.001	0.079	NA	NA
	Cop. GAN	101/142	1.14	0.017	0.172	NA	NA
	CTGAN	115/142	1.10	0.027	0.151	NA	NA
Acute Myeloid Leukemia	Simulants	0/108	0.19	0.000	0.081	0.06	NA
	GC GAN	36/108	0.29	0.007	0.192	0.68	NA
	TVAE	58/108	0.30	0.066	0.387	5.86E-35	NA
	Cop. GAN	44/108	0.42	0.032	0.304	0.11	NA
	CTGAN	36/108	0.42	0.021	0.311	0.20	NA

Takeaways

- **Simulants outperforms all other approaches in fidelity benchmarks** across multiple different clinical trial datasets (note: each dataset composed of 3-20 clinical trials)

1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis** - AMIA 2022

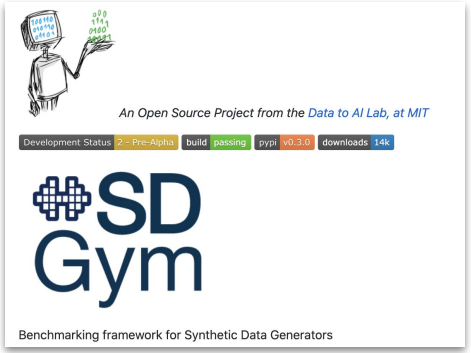
Results – Fidelity & cross validation

Benchmarks – Simulants outperforms all methods in canonical data too

Outperforms state-of-the-art GANs, VAEs and RNNs in canonical datasets at all scales in open benchmarks (SD-GYM / MIT AI Lab)

Group	Synthesizer	Description	Real Datasets: mnist-12 (60000x145)	mnist-28 (60000x785)	news (31644x59)	adult (22561x15)	covtype (481012x55)	credit (264807x30)	intrusion (394021x41)	census (199523x41)
Embedding	Simulants	Our method which uses the high dimensional embeddings.	0.9	0.9	0.1	0.8	0.8	1	1	0.9
Trivial	Identity	The synthetic data is the same as training data.	0.9	0.9	0.1	0.8	0.8	1	1	0.9
	Uniform	Each column in the synthetic data is sampled independently and uniformly.	0.1	0.1	-4	0.5	0.1	0.6	0.1	0.5
	Independent	Each column sampled independently. Continuous columns use Gaussian Mixture Model and discrete columns use the PMF of training data.	0.1	0.2	-0.06	0.6	0.4	0.9	0.7	0.7
GANs	Medgan	Minibatch averaging to efficiently avoid mode collapse.	0.4	0.1	-6	0.6	0.4	0.9	0.9	0.6
	VEEGAN	The method features a reconstructor network, reversing the action of the generator by mapping from data to noise.	0.4	0.2	-3.00E+08	0.7	0.2	0.9	0.5	0.8
	CTGAN	Models discrete and continuous columns.	0.1	0.1	-0.07	0.8	0.6	1	1	0.9
	CopulaGAN	Uses GaussianCopulas to make the underlying CTGAN model task of learning the data.	0.2	0.2	-0.06	0.8	0.6	1	1	0.9
	TableGAN	Ggenerates synthetic data using a convolutional neural network which optimizes the label column's quality.	0.1	0.1	-6	0.8	0	1	nan	0.9
VAEs	TVAE	Based on the VAE-based Deep Learning data synthesizer.	0.8	0.8	-0.02	0.8	0.7	1	1	0.9
Others	CLBN	Uses Bayesian networks.	0.7	0.2	-7	0.8	0.6	1	0.9	0.9
	PrivBN	A differential privacy method which uses a Bayesian network to model the correlation among the attributes.	nan	nan	nan	0.8	0.5	1	0.9	0.9
	GaussianCopulaCategorical	Based on copula functions and uses a CategoricalTransformer.	0.1	nan	-5		0.4	1	1	0.9
	GaussianCopulaCategoricalFuzzy	Based on copula functions and uses a CategoricalTransformer with fuzzy=True.	0.2	0.2	-9	0.8	0.4	1	0.8	0.8
	GaussianCopulaOneHot	Based on copula functions and uses a OneHotEncodingTransformer.	0.5	0.5	-4	0.8	0.5	1	0.9	0.9

Legend: Best match / Worst match



MEDIDATA

1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis** - AMIA 2022

DASSAULT SYSTEMES

Designing a privacy framework for synthetic clinical trial generation

MEDIDATA

Icons Source: Flaticon; Full credits in the last slide

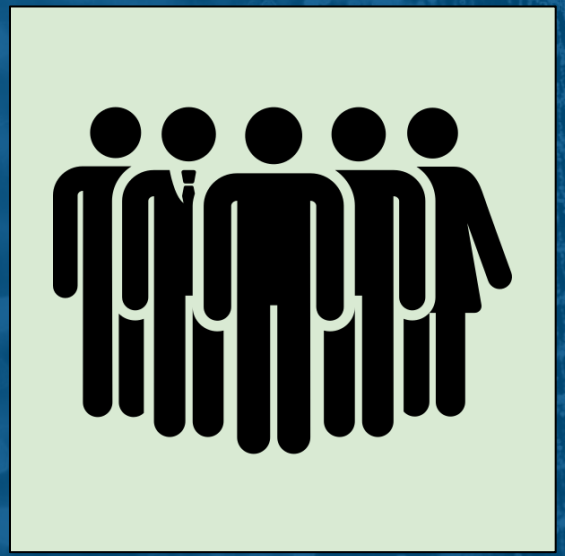
DASSAULT SYSTEMES

Synthetic data is a solution to encourage collaboration and continue innovation in pharma

Sponsor A

Sponsor B

Sponsor C



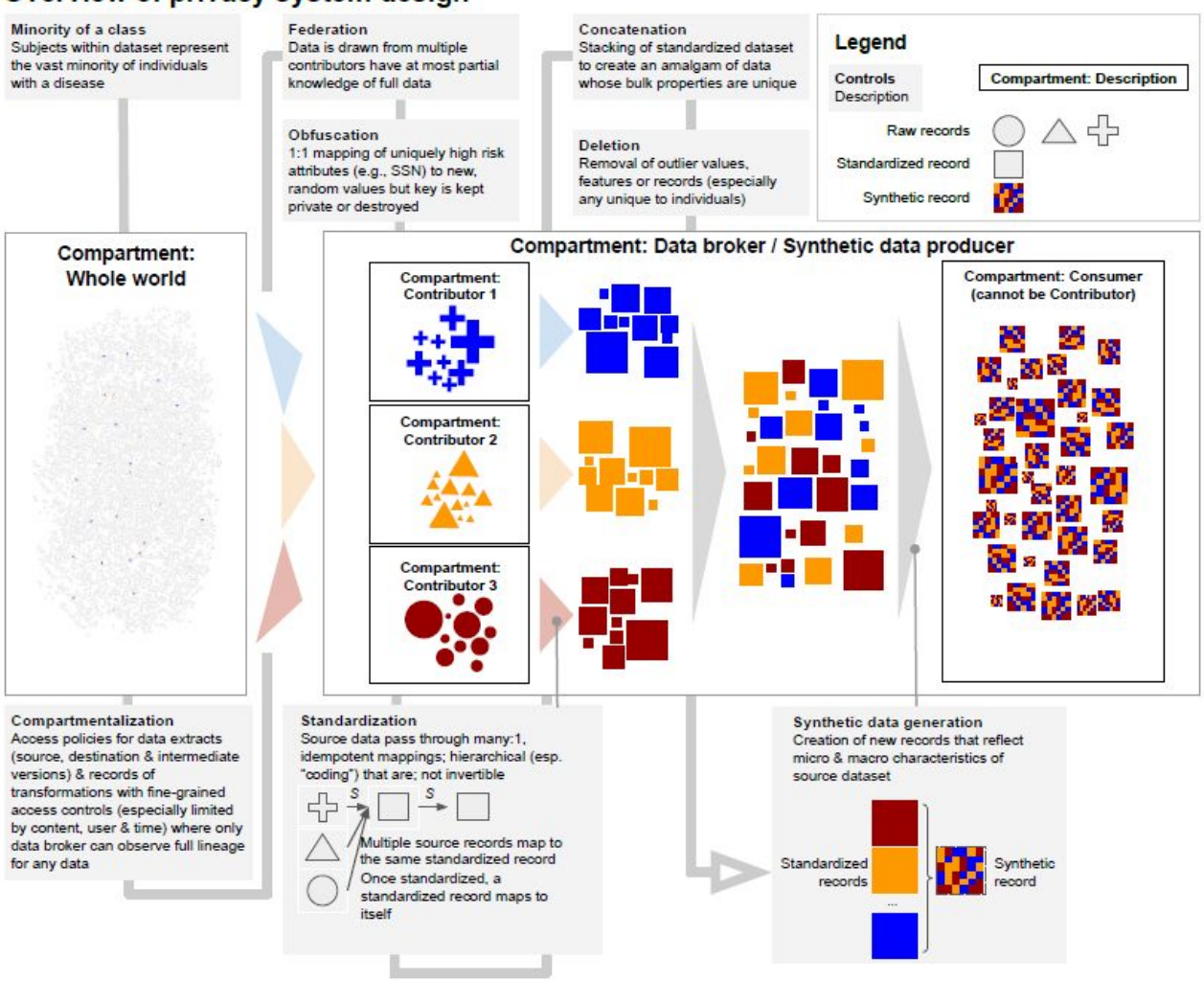
MEDIDATA

Icons Source: Flaticon; Full credits in the last slide

DASSAULT SYSTEMES

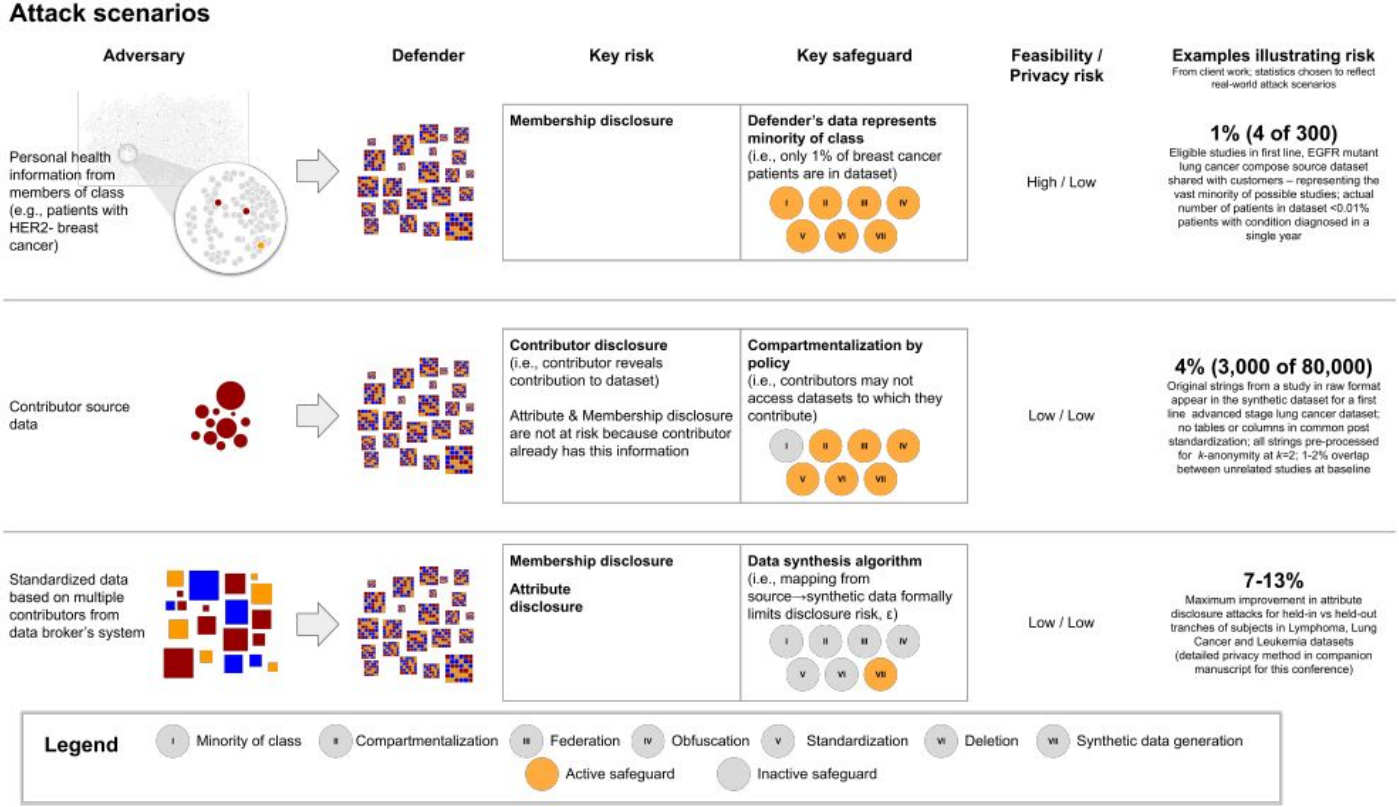
Personas	Competing interests			
	Data Contributor privacy	Patient privacy	Privacy of proprietary information	Synthetic data fidelity
 Data contributor	✔	✔	✔	
 Synthetic Data broker	✔	✔	✔	✔
 Data consumer				✔

Icons Source: Flaticon; Full credits in the last slide



1. A Shafquat, J Mezey, M Beigi, J Sun, JW Aptekar A source data privacy framework for synthetic clinical trial data- NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research

Adversarial Scenarios



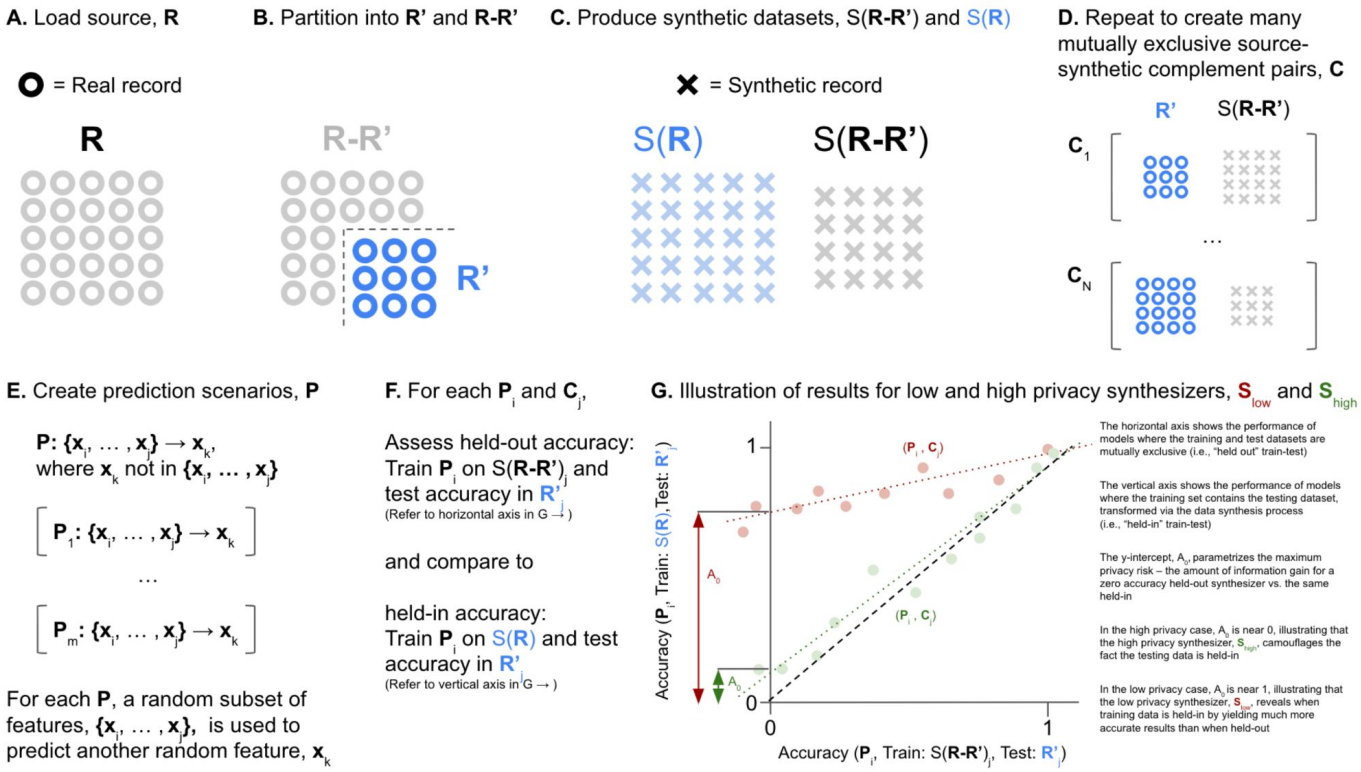
1. A Shafquat, J Mezey, M Beigi, J Sun, JW Aptekar A source data privacy framework for synthetic clinical trial data- NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research

MEDIDATA

NEURAL INFORMATION PROCESSING SYSTEMS

DASSAULT SYSTEMES

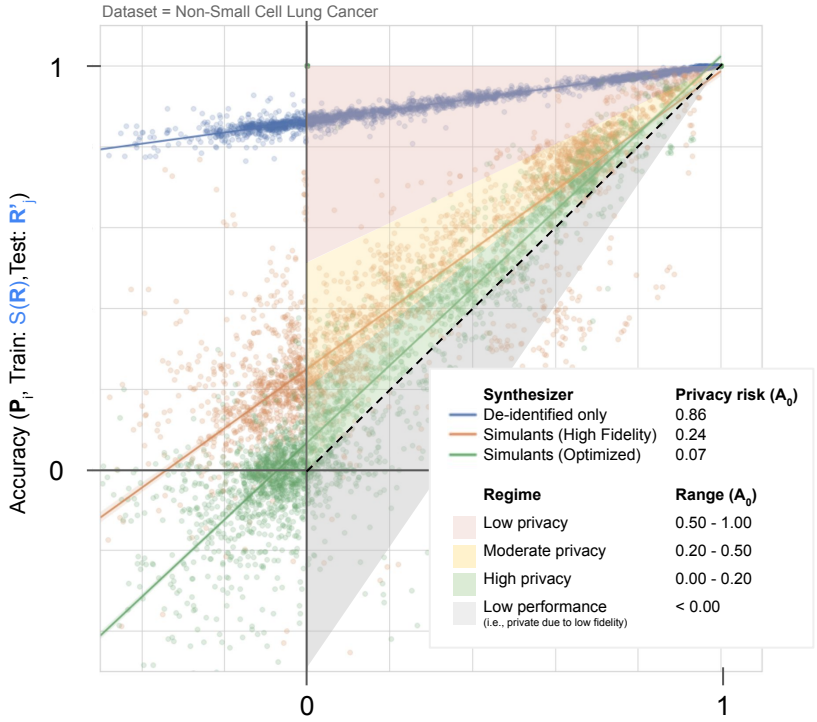
Overall privacy preservation score



1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis** - AMIA 2022

Trial simulation- Standard tests, privacy

Representative privacy plots for high-fidelity & optimized Simulants

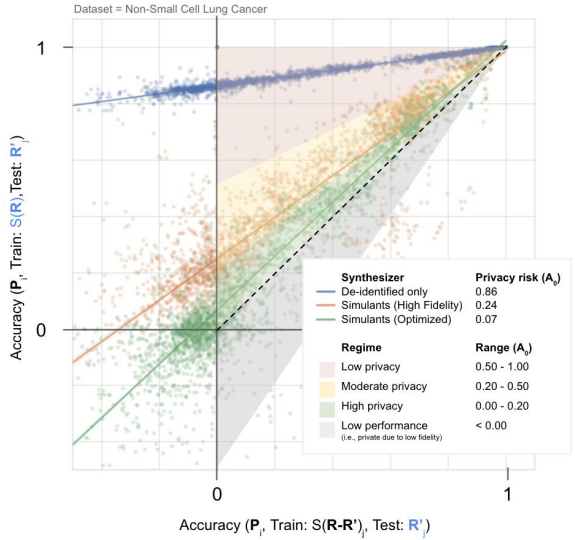


1. M Beigi, A Shafquat, J Mezey, JW Aptekar Synthetic Clinical Trial Data while Preserving Subject-Level Privacy - NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research
2. M Beigi, A Shafquat, J Mezey, JW Aptekar **Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis** - AMIA 2022

Results – Privacy

Benchmarks

H. Representative privacy plots for high-fidelity & optimized Simulants



I. Quantification of privacy risk (A_0) by synthesizer over datasets

Synthesizer	Dataset: NSCLC	DLBCL	AML	ALL
De-identified only	0.86	0.86	0.77	0.86
Simulants (Optimized)	0.07	0.09	0.13	0.08
Gaussian Copula GAN	0.05	0.05	0.04	0.02
TVAE	0.08	0.26	0.05	0.10
Copula GAN	-0.14	-0.01	-0.06	-0.14
CTGAN	-0.18	-0.12	-0.10	-0.10

Takeaways

- **Simulants delivers high privacy** at par with all other synthesizers tested (including Generative Adversarial Networks [GANs] and Variational Autoencoders [VAEs])
- Results for Novartis project are consistent with these benchmarks

MEDIDATA

DASSAULT SYSTEMES

Data Augmentation and applications in machine learning

Icons Source: Flaticon; Full credits in the last slide

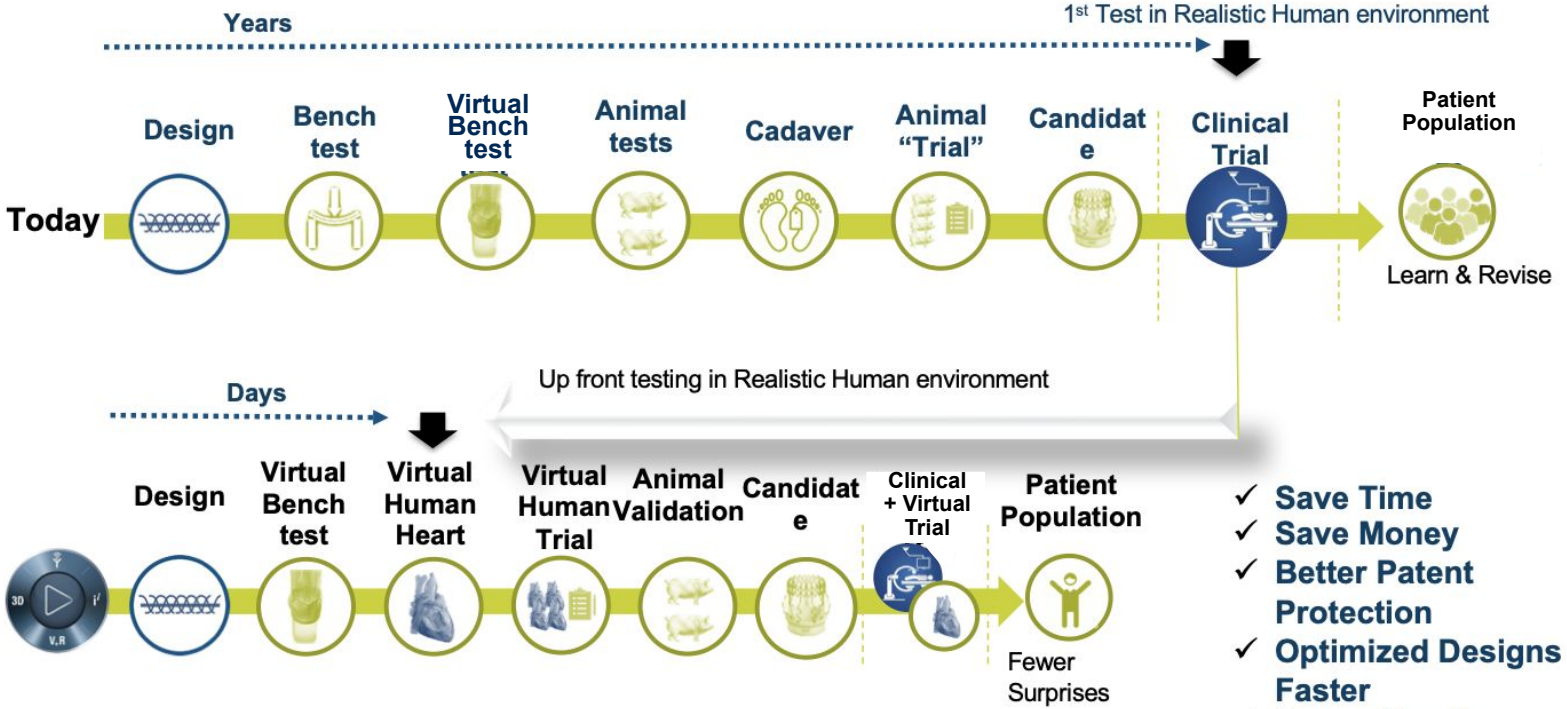
Accelerating *in silico* clinical trials

Full presentation available here:
https://events.3ds.com/living-heart-and-virtual-twin-for-humans-symposium

MEDIDATA

DASSAULT SYSTEMES

Value Proposition for *In Silico* Clinical Trial – "Years to Days"



60% of cost is people on project – Time is Money!

- ✓ Save Time
- ✓ Save Money
- ✓ Better Patent Protection
- ✓ Optimized Designs Faster
- ✓ Fewer Recalls
- ✓ Physician training

MEDIDATA

DASSAULT SYSTEMES

iSCT Key Challenge – Defining & Proving the Hypothesis

Identify strong correlation between acute model outputs (surrogate endpoints) & clinical endpoints

Clinical Outcomes = F (Clip Performance X Complexity of Anatomy X

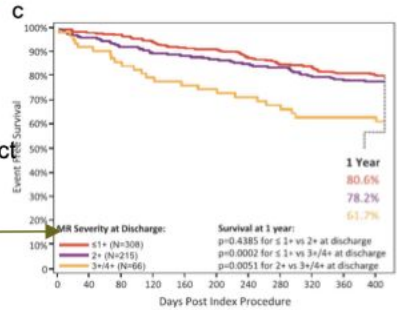
Experience)

X Physiological Adaptation X Clinical Factors* X Operator)

???

Primary device function

Secondary health impact



"acute model outputs" ↔ "observed clinical outcome"

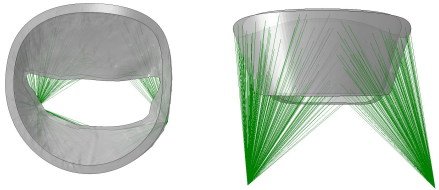
Correlation?

**Circulation. 2019;139:37–47. DOI: 10.1161/CIRCULATIONAHA.117.031733

Virtual Patient Engine Schematic

Input:

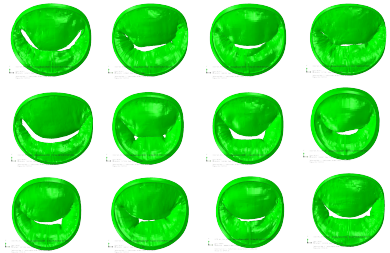
Initial VPC - A collection of (physics-based) patient model definitions & pre-operative simulation results



Top view Front view

Output:

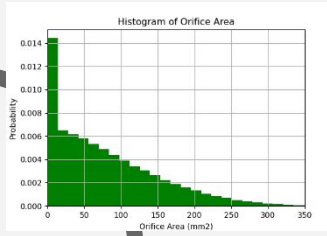
iSCT VPC - A physics-based VPC with targeted pre-operative characteristics to be treated in an iSCT



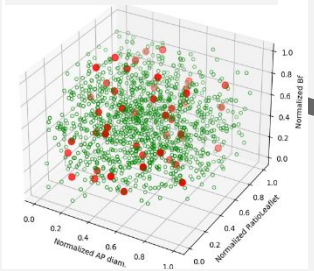
Machine Learning Powered VPE

Create surrogate model to accelerate VPC creation

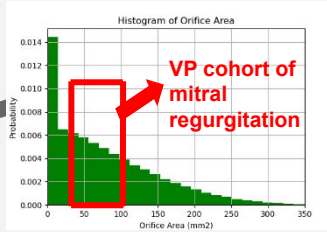
Build surrogate-based VPC



Obtain targeted representative VPs



Down sample to get VPs with user defined characteristics



MEDIDATA

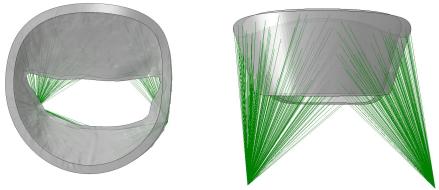
Slide copied from the SIG 35 readout:
https://www.3ds.com/products-services/simulia/solutions/life-sciences-healthcare/the-living-heart-project/

DASSAULT SYSTEMES

Virtual Patient Engine Schematic

Input:

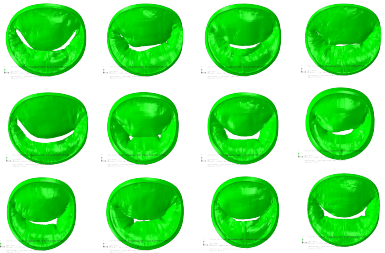
Initial VPC - A collection of (physics-based) patient model definitions & pre-operative simulation results



Top view Front view

Output:

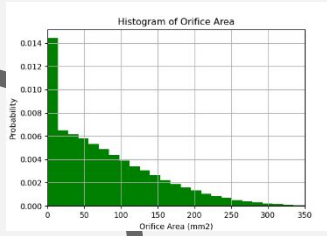
iSCT VPC - A physics-based VPC with targeted pre-operative characteristics to be treated in an iSCT



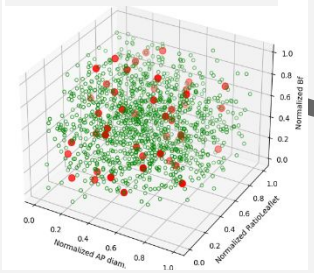
Machine Learning Powered VPE

Create surrogate model to accelerate VPC creation

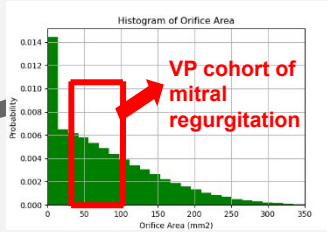
Build surrogate-based VPC



Obtain targeted representative VPs



Down sample to get VPs with user defined characteristics

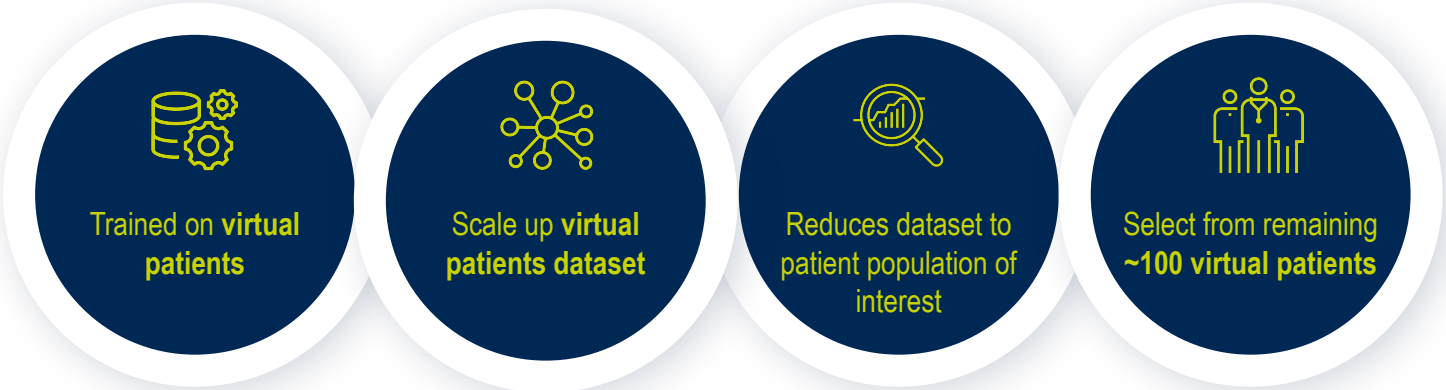


MEDIDATA

Slide copied from the SIG 35 readout:
https://www.3ds.com/products-services/simulia/solutions/life-sciences-healthcare/the-living-heart-project/

DASSAULT SYSTEMES

Virtual Patient Engine Schematic



Create surrogate model to accelerate VPC creation

Build surrogate-based VPC

Down sample to get VPs with user defined characteristics

Obtain targeted representative VPs

Computationally expensive and impractical to scale

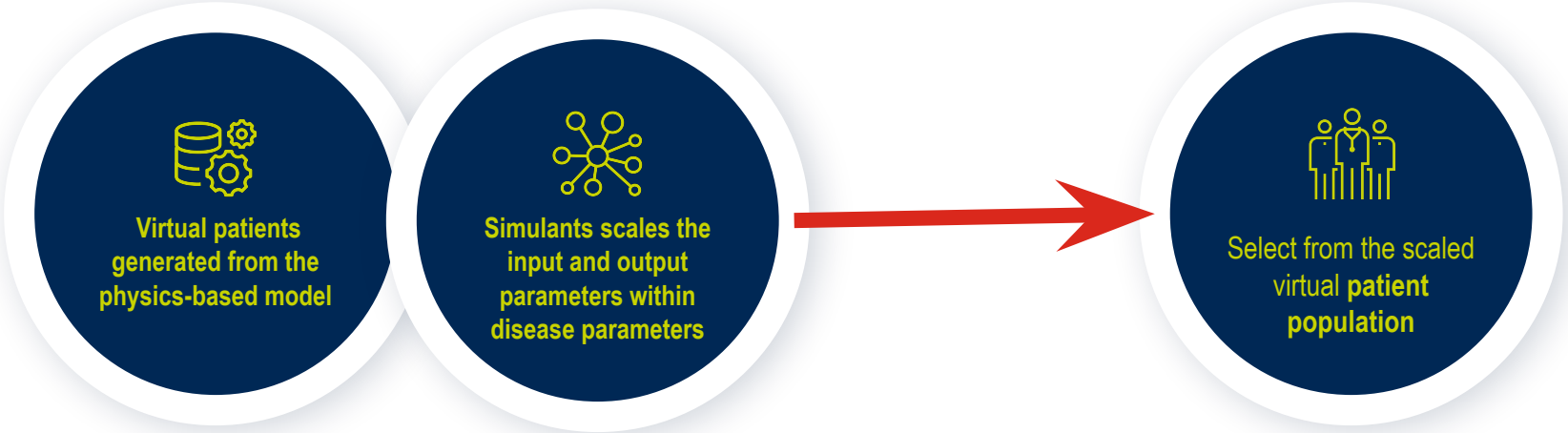
Generates redundant samples that don't pass the quality control criteria

14x reduction in total sample size for selection

MEDIDATA

DASSAULT SYSTEMES

Virtual Patient Engine Schematic



Training dataset generated using the initial physics-based population

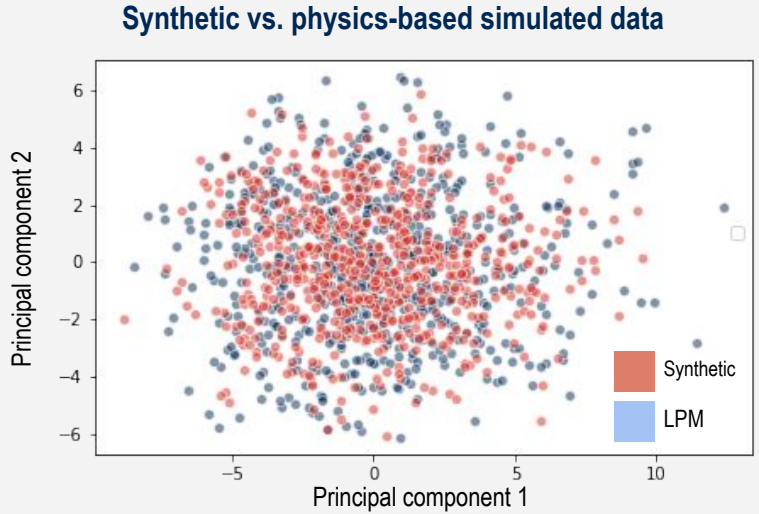
Simulants pipeline generates synthetic data comprising of parameters of interest

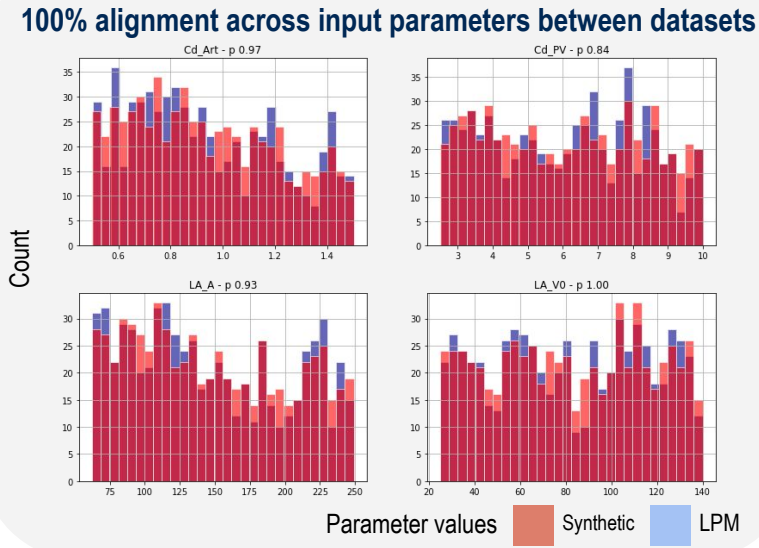
Obtain targeted representative VPs

MEDIDATA

DASSAULT SYSTEMES

Synthetic data aligns with physics-based model simulated patients





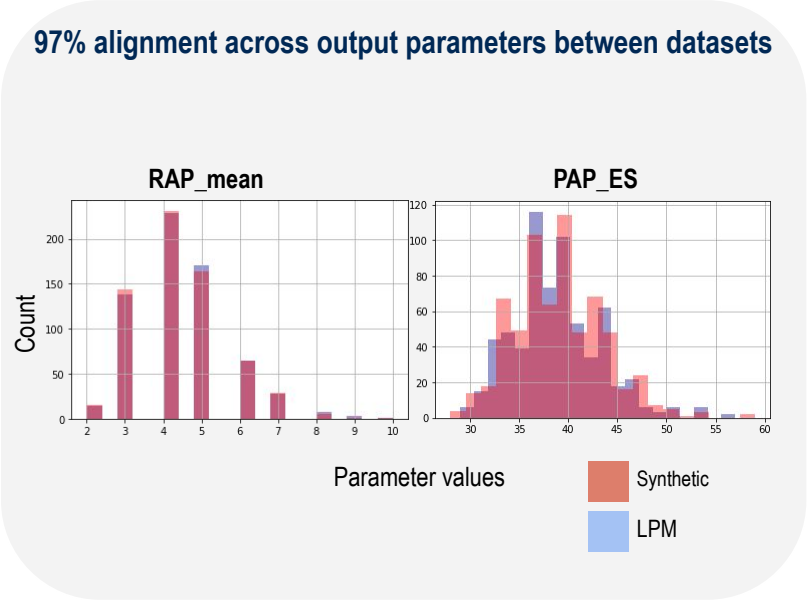
Alignment between source/template and Simulants generated synthetic dataset indicates synthesis reflects the source population well

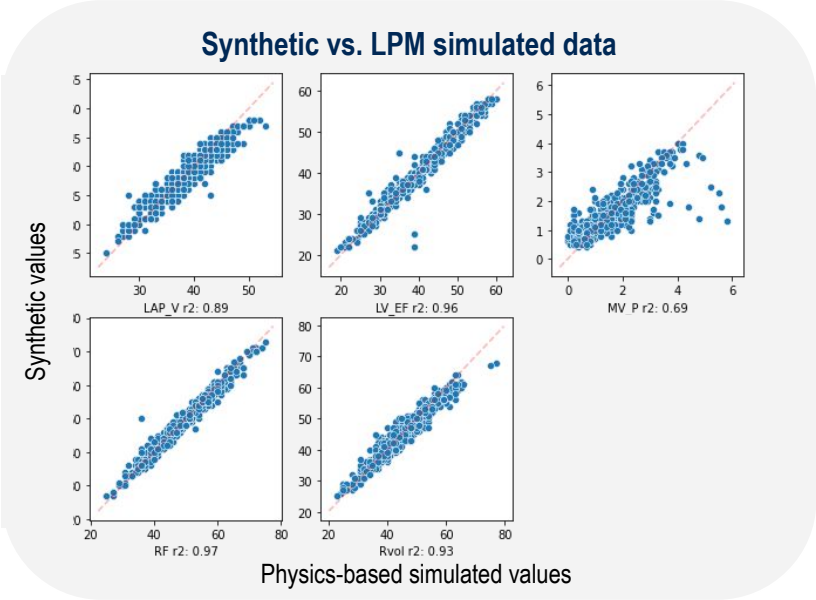
Alignment between input parameters is evaluated by performing the K-S test for input parameters produced by Simulants and the physics-based model. All 14 input parameters had non-statistically significant differences using the K-S test.

MEDIDATA

ILT ES

Synthetic data aligns with physics-based model simulated patients

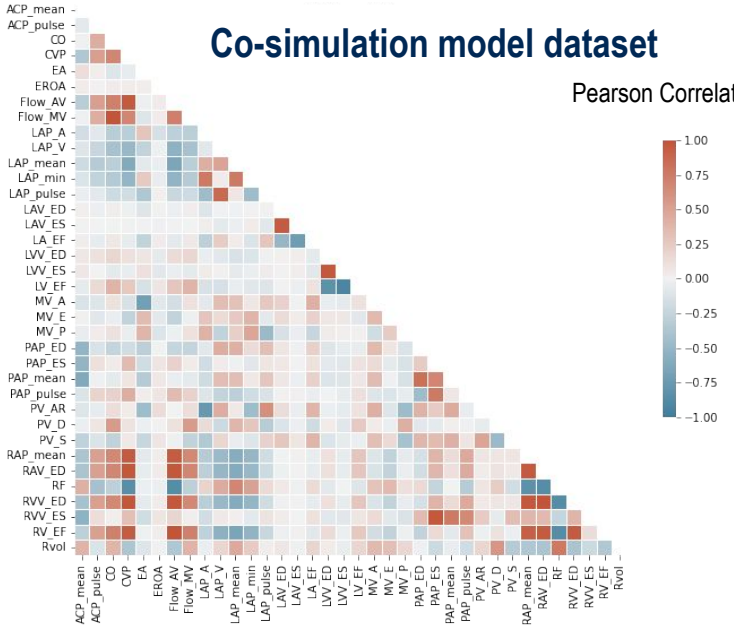


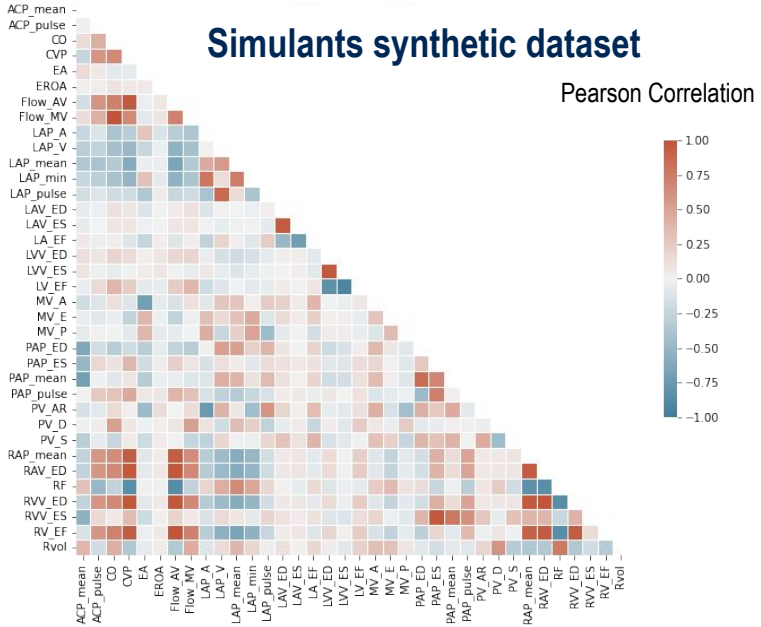


Alignment between output parameters is evaluated by performing the K-S test for output parameters produced by Simulants and the physics-based model. 34 parameters out of 35 parameters (i.e. MV_P) had non-statistically significant differences.

The diagonal in each plot indicates model performance on par with expectation. The x-axis shows the parameter values from the physics-based model and y-axis shows the Simulants-generated parameter values. High r^2 values (close to 1) indicate Simulants performance and agreement between the two models

Preservation of bidirectional relationships in synthetic data





Heatmaps indicate correlation across features in the source/physics-based simulated dataset and Simulants synthetic dataset. The similarity between the heatmaps indicates the bidirectional relationships and correlations observed in the dataset generated using the physics-based patient model are preserved in the Simulants-generated synthetic dataset.

Acceleration via synthetic data

97% Agreement between simulated and synthetic output parameter distributions

100% Agreement between simulated and synthetic input parameter distribution

100% Preservation of sample size

Simulants is successfully able to mimic the the physics-based co-simulation model

DASSAULT SYSTEMES

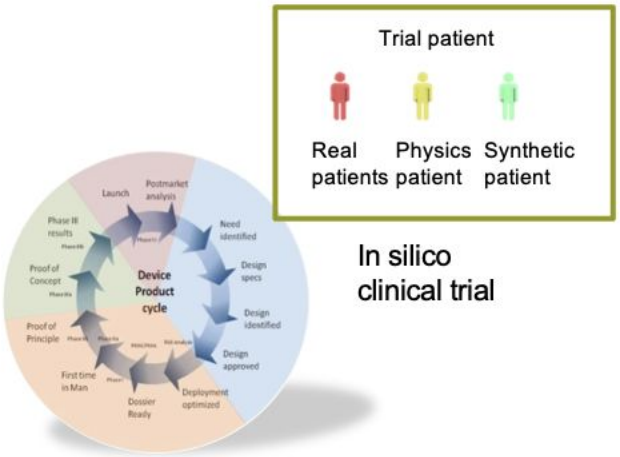
Summary of advantage of Simulants

Though the current VPE framework offers a competitive advantage over the Finite Element Analysis to simulate the living heart, a Simulants powered pipeline offers:

- **Fast and robust** way to generate synthetic parameter sets that reflect the physics-based parameter simulation
- **Reduce the computational cost** of running the Finite Element Analysis where the current process is slow and intensely computationally expensive
- **Doesn't depend on domain knowledge** of parameter boundary values
- **Preservation and scaling of sample size** of the physics-based dataset
- **Removal of redundancy** in the pipeline by only producing parameter sets that are aligned with the parameter distribution
- Potential to **target synthetic generation** of the distribution of interest (e.g. mitral valve regurgitation)

Future directions

Conduction 'in Silico clinical trials' using patient-specific models, physics based population models and synthetic patients to form virtual cohorts for testing the safety and/or efficacy of new drugs and of new medical devices.



In silico clinical trial

Real patient (Patient specific models, acute and long end points, clinical outcome)

Physics patient (population models, detailed physics and physiology correlations, defining hypothesis)

Synthetic patient (Enrichment with VP augmentation, identifying correlation between acute and clinical outcome, proving hypothesis)

Key takeaways

Data sharing is critical for continued innovation in the biotech and pharma industry

Synthetic data provides a fast, secure and reliable way to share private data

Synthetic data allows an innovative way to synthesize and augment training datasets to improve AI/ML model performance

Icons Source: Flaticon; Full credits in the last slide

Presentation credit:

Jacob Aptekar, MD PhD
VP, Medidata AI

Mandis Beigi, PhD
Sr. Director, Medidata AI

Jia Chen
Senior Director, Medidata AI

Afrah Shafquat

Sr. Data Scientist II, Medidata AI

Afrah.Shafquat@3ds.com

afrahshafquat

MEDIDATA

DASSAULT SYSTEMES

Credits

1. Neural-network icons created by Freepik - Flaticon
2. Corporate icons created by Sumitsaengtong - Flaticon
3. Consumer icons created by Freepik - Flaticon
4. Consumer icons created by Freepik - Flaticon
5.
6. Sheet icons created by Andrejs Kirma - Flaticon
7. Neural network icons created by Freepik - Flaticon
8. Patient icons created by SBTS2018 - Flaticon
9. Privacy icons created by Fathema Khanom - Flaticon
10. Clinical icons created by Parzival' 1997 - Flaticon
11. Collaboration icons created by Freepik - Flaticon
12. Collaboration icons created by Freepik - Flaticon
13. Group icons created by Freepik - Flaticon
14. Group icons created by Freepik - Flaticon
15. Denied icons created by Alfredo Creates - Flaticon
16. Patient icons created by SBTS2018 - Flaticonk icons created by Freepik - Flaticon
17. Privacy icons created by Freepik - Flaticon
18. Collaboration icons created by small.smiles - Flaticon
19. Collaboration icons created by small.smiles - Flaticon
20. Medicine icons created by max.icons - Flaticon
21. Simulation icons created by Freepik - Flaticon