Simulants: Unlock insights and accelerate medical breakthroughs with synthetic clinical trial data

Overview:

Every day, data from clinical trials is entered into electronic data capture (EDC) systems in support of finding the next medical breakthrough. This data – complete with covariates, endpoints, and health data captured as part of the clinical trial protocol – provide rich insights into the safety and efficacy of new therapeutics, but remain locked away within individual organizations and systems rendering them largely inaccessible.

Utilizing historical clinical trial data can unlock and accelerate research and development by providing insights into what has already worked, and what hasn't. However, as with other sensitive data containing patient identifiable information (PII) or intellectual property (IP), access to clinical trial data is limited by regulatory requirements, technical protection protocols, high proprietary value of clinical trial data, and strict privacy requirements required by sponsors stemming from issues of patient consent and preservation of patient trust. These barriers are significant enough that even with clinical data sharing commitments, policies, and protocols, sharing of de-identified patient level clinical trial data remains vanishingly rare.

The emergence of generative AI now enables the creation of synthetic data, data generated to mimic the properties and patterns in the original data source while safeguarding private details. Synthetic data is a type of data that has been generated artificially or algorithmically, as opposed to data that is collected from real people or events, therefore offering a revolutionary new way to share and access the most sensitive, yet often crucial, data.

Key-takeaways:

Synthetic data offers a practical, efficient, and secure way to access sensitive data by preserving the characteristics of an underlying dataset while safeguarding privacy

Simulants is a first-of-its-kind synthetic data solution that generates synthetic data from cross-sponsor historical clinical trial data, complete with the covariates, endpoints, and variables as captured through the clinical trial protocols– ensuring a high fidelity dataset

Clinical developers can now leverage Simulants to (i) unlock crucial insights contained with historical trial data and (ii) address bias within current trial data through the creation of balanced, representative synthetic clinical trial data.

Why use synthetic data?

Synthetic trial data can offer the unique ability for clinical developers and researchers to interact with cross-sponsor historical clinical trial data, while protecting patient privacy and IP.

Preserve patient privacy

Patients' explicit consent for research is respected when using synthetic data, promoting trust and ethical data handling in the medical research community.

Protect sponsor IP

Synthetic data generation offers a practical and accessible solution for researchers, allowing them to work with high-quality data without the need for extensive data-sharing agreements or complex data handling procedures.

Accelerate medical breakthroughs

Make the drug development process more efficient by using data to design better trials, optimize inclusion/exclusion criteria, and identify relevant endpoints from past trials.

Simulants:

Simulants is a groundbreaking synthetic data product – uniquely derived from Medidata's exclusive repository of global, standardized historical clinical trial data, containing data over 30,000 cross-sponsor trials and 9 million patients. This comprehensive underlying dataset includes covariates, endpoints, and variables as captured in their respective protocols. Powered by Medidata's patented algorithm, Simulants is optimized for generating high-fidelity synthetic data from clinical trial datasets, which contain uniquely sensitive data compared to other typical datasets used for synthetic data. Proven to meet rigorous regulatory and privacy constraints, Simulants empowers clinical developers with expanded data access, leading to increased research accuracy, insightful evidence, and optimized clinical trial design.

Use cases unlocked by Simulants:

Deepen disease understanding

- Optimize trial protocols and study parameters
- Identify subpopulations with the highest unmet medical need
- Construct a rich patient journey and compare treatment outcomes over time and across populations; evaluate real world treatment patterns
- Identify biomarker correlation with treatment efficacy

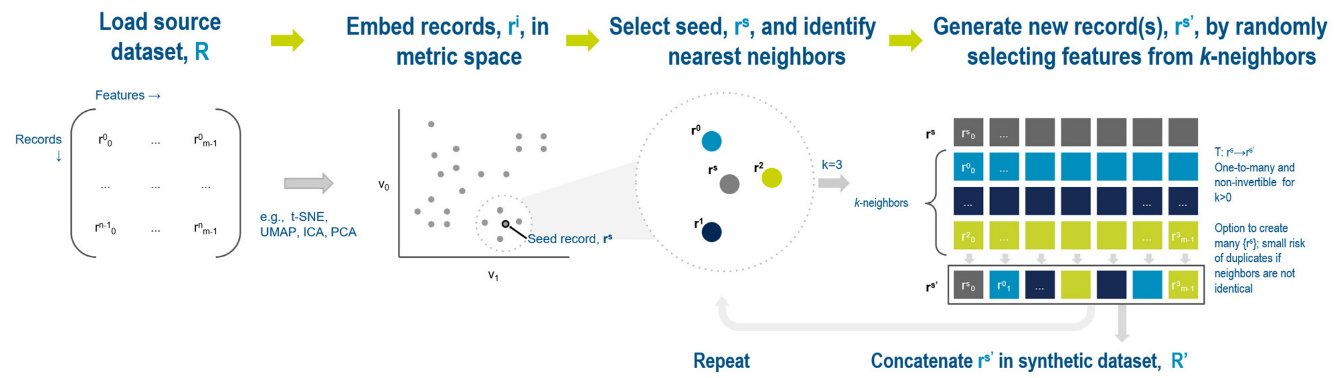
Increase participant safety

- Identify patients most likely to drop-out and develop strategies to engage or recruit
- Identify patients most at risk for a serious adverse event
- Identify early indications of serious adverse event and monitor long term safety outcomes

Increase the probability of trial success

- Model different trial scenarios to design more effective trials
- Use synthetic data to train or validate models, including the ability to create a more balanced dataset and "upsample" underrepresented groups within your data
- Generative comparative evidence for clinical development decision-making, regulatory submissions or payor conversations
- Identify "impossible" to recruit patient populations

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

Repeat

Concatenate $r^{s'}$ in synthetic dataset, R'

Features →
Records ↓
r^0_0 … r^0_{m-1}
… … …
r^{n-1}_0 … r^n_{m-1}
e.g., t-SNE, UMAP, ICA, PCA
v_0
v_1
Seed record, r^s
r^0
r^s
r^2
r^1
k=3
k-neighbors
r^s
r^s_0 …
r^0_0 …
…
… …
r^2_0 … … r^3_{m-1}
$r^{s'}$
r^s_0 r^0_1 … r^3_{m-1}
T: $r^s \rightarrow r^{s'}$
One-to-many and non-invertible for k>0
Option to create many {r^s}; small risk of duplicates if neighbors are not identical

Publications

US Patent: System and method for generating a synthetic dataset from an original dataset

ICML 2023: An interpretable data augmentation framework for improving generative modeling of synthetic clinical trial data (Best paper award)

Neurips 2022: A source data privacy framework for synthetic clinical trial data

Neurips 2022: Synthetic Clinical Trial Data while Preserving Subject-Level Privacy

AMIA 2022: Simulants: Synthetic Clinical Trial Data via Subject-Level Privacy-Preserving Synthesis

Medidata, a Dassault Systèmes company, is leading the digital transformation of life sciences.
Discover more at **www.medidata.com** and follow us **@medidata**. Contact us at **info@medidata.com** | **+1 866 515 6044**