Paper TT02

Why Wait for Real Data? The Power of Synthetic Data in Modern Clinical Trials

Abhijeet Arora, GenInvo, Toronto, Canada
Inderjeet Arora, GenInvo, Toronto, Canada

ABSTRACT

Clinical trials often struggle to access diverse, high-quality data due to delays in data collection, privacy restrictions, and recruitment biases. These challenges slow down study progress and make it harder to build and validate statistical programs early in the process. Synthetic data generation presents a powerful alternative by producing realistic, privacy-preserving datasets that closely mimic real clinical trial data. With open-source Python libraries or advanced platforms, teams can create synthetic datasets to develop and test edit checks, analysis scripts, and models well before real data becomes available. This approach accelerates development, minimizes rework, and reduces dependency on sensitive patient information. Beyond improving efficiency, synthetic data supports more inclusive and unbiased analyses by simulating a broader range of patient profiles. Ultimately, integrating synthetic data with AI enables faster, safer, and more reproducible clinical research—helping organizations meet regulatory expectations while driving innovation across the drug development lifecycle.

INTRODUCTION

Clinical trials lie at the heart of bringing novel medical therapies to patients. Yet the path from protocol to submission is fraught with delays, inefficiencies, and data challenges. Among the most persistent bottlenecks are:

1. **Data delays** — collecting, cleaning, and validating actual patient data takes months or years.
2. **Privacy and regulatory constraints** — direct access to patient-level data is tightly controlled, complicating early-stage development.
3. **Bias and lack of diversity** — real-world recruitment frequently underrepresents certain demographics, limiting generalizability.
4. **Rework risk** — programming pipelines are often built "blind," with the real data arriving late, causing rework in deliverables.

Given these realities, the question arises: must we always "wait for the data" before building robust, end-to-end deliverables (edit checks, SDTM/ADaM pipelines, TFLs)? Or can we simulate a credible substitute in advance?

Synthetic data promises a middle path: generate realistic, privacy-safe datasets early, and use them as a sandbox to build, test, and refine clinical trial modules—well in advance of real data arrival. This approach offers an opportunity to front-load validation, detect errors early, and mitigate risks before "go-live."

In this paper, we synthesize recent advances in synthetic data methods, illustrate applications in the clinical trial domain, and propose a conceptual workflow using open-source and commercial tools. Our aim is to present a compelling, technically rigorous argument to sponsors, data scientists, and regulators: synthetic data is not a gimmick—it is a bridge to more efficient, robust, and trustworthy trial execution.

CURRENT CHALLENGES IN CLINICAL TRIALS

DATA DELAYS AND PIPELINE FRAGILITY

In many trials, programming deliverables (e.g. SDTM generation, edit checks, TFL skeletons) must be templated before real data is available. When data arrives, unexpected anomalies or edge cases often surface, forcing rework, debugging, and delay. The blind zone between design and data maturity is a chronic source of inefficiency.

PRIVACY, CONFIDENTIALITY & ACCESS BARRIERS

Accessing real patient-level data involves strict governance (e.g. data use agreements, de-identification procedures, audit trails). These constraints limit early-stage experimentation, especially for external partners or cross-functional teams. Synthetic data, by definition, contains no actual patient identifiers and thus bypasses many privacy constraints.

SELECTION BIAS AND REPRESENTATIVENESS

Clinical trial populations often skew (by geography, demographics, disease subtype, or site selection). Underrepresented subgroups weaken inference and robustness. Synthetic data can be designed to simulate broader, more balanced populations—offering a "stress test" for pipelines under more challenging distributions.

REGULATORY TRUST & VALIDATION RISK

Regulators and auditors are cautious about black-box methods or simulated data. There is a legitimate concern that synthetic datasets may mislead downstream statistical analyses or induce biases. Therefore, any synthetic data approach must be transparent, validated, and accompanied by metrics of fidelity, privacy, and utility.

This is not speculative: a recent systematic review highlighted the challenge of evaluating synthetic tabular data—not just for resemblance but also for utility and privacy tradeoffs. Another survey of synthetic longitudinal data methods noted that very few approaches embed privacy mechanisms, and most evaluations ignore privacy altogether.

Therefore, to gain adoption in regulated contexts, synthetic data workflows must address not only realism and utility, but also explainability, validation, and governance.

SYNTHETIC DATA OVERVIEW

DEFINITION & OBJECTIVES

"Synthetic data" refers to data artificially generated (via models, simulations, or transformations) to mimic the statistical and logical properties of real data without containing actual individual-level records. It is not anonymized real data—it is entirely novel, yet constructed to preserve relationships, distributions, and constraints relevant to downstream use.

We might frame the objectives of synthetic clinical data generation as:

- **Resemblance** — the synthetic data should mirror marginal and joint distributions, time relationships, correlation structures, and clinical logic.
- **Utility** — synthetic data must support intended downstream workflows (e.g. edit checking, modeling, TFL development) with minimal divergence from real-data performance.
- **Privacy safety** — synthetic data should not permit re-identification or inference of real patients, typically enforced by privacy-preserving techniques (e.g. differential privacy, noise injection, k-anonymity).

GENERATION TECHNIQUES

Several broad categories of synthetic data generation dominate:

1. **Parametric / rule-based simulation:** Based on statistical models (e.g. multivariate distributions, copulas) and domain rules (e.g. visit windows, lab ranges). Offers interpretability and control, but may lack complexity for real-world heterogeneity.

2. **Generative models / machine learning:** Models such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), or Bayesian networks can learn complex joint distributions. But they can suffer from mode collapse, overfitting, or privacy leakage if not properly constrained. A benchmarking study of synthetic EHR models emphasized that no single model dominates across all utility/privacy dimensions.

3. **Hybrid / constraint-based generative:** Combine rule-based logic (e.g. visit ordering, clinical rules) with learned residual distributions. This balances realism with interpretability.

4. **Post hoc perturbation / anonymization of real data (less true synthetic):** Methods that start with real records and add noise or shuffle attributes. While useful for certain tasks, such approaches risk privacy leakage and may not suit early-stage pipeline development.

Selecting a method requires matching the intended use case, dataset complexity (e.g. longitudinal, multilevel), and required privacy guarantees.

ETHICAL & REGULATORY CONSIDERATIONS

- **Trust and explainability**: Stakeholders demand transparency in how synthetic data was generated (e.g. which constraints, noise injection strategies, model choice).
- **Validation and metrics**: Common metrics include statistical distance, predictive model performance, coverage of edge cases, and privacy risk scores.
- **Governance**: Synthetic data should be versioned, documented, and audited, with metadata tracking and provenance.
- **Regulatory acceptance**: Synthetic data is not yet considered a substitute for actual analysis in submissions, but its use to support development, QC, and risk mitigation is gaining traction.

A recent position paper argues that building trust is essential: synthetic data systems must demonstrate consistent performance, transparent assumptions, and domain-aligned validation to win adoption in clinical AI workflows.

Thus, synthetic data must be viewed not as a “magic fix” but as a disciplined engineering and governance exercise.

APPLICATIONS IN CLINICAL TRIALS

Synthetic data can support multiple upstream and midstream processes in clinical trial development, particularly in a CDISC-centric environment.

EDIT CHECK AND VALIDATION TESTING

Using synthetic SDTM-like datasets, you can stress-test your edit checks (e.g. range checks, cross-variable consistency, chronological checks). You can deliberately inject edge cases or rare combinations to validate robustness. Solutions can explicitly be designed for this: generating “meaningful synthetic data” that adheres to expected controlled terminology, date ordering, and ability to introduce outliers or anomalous values.

PROGRAMMING TEMPLATE & SHELL BUILD

With synthetic data in hand, programmers can build SDTM mapping logic, define derived variables, and prototype ADaM derivations and mock TFL skeletons. This front-loads debugging and accelerates the hand-off once real data arrives.

QA / QA DRY RUNS

Dry-run of the full pipeline—mapping, derivations, integrity checks, TFL generation—on synthetic datasets helps expose hidden bugs, efficiency bottlenecks, or logic gaps before real data is deployed. This acts as a rehearsal, reducing surprises.

MODEL DEVELOPMENT & AI/ML WORKFLOWS

In settings where machine learning or predictive modeling is embedded, synthetic data allows early feature engineering, model prototyping, or training, without risking confidentiality. Because you can generate large volumes, you can explore rare-event sampling or stress-test model behavior under distribution shifts (e.g. varying demographics).

SENSITIVITY / SCENARIO MODELING

You can simulate “what-if” populations—e.g. adjusting prevalence of a comorbidity, shifting demographics, or trial dropout—to test robustness of statistical programs and TFL sensitivity.

Example: External / Synthetic Control Arms

While synthetic arms are still emerging in regulatory practice, combining real controls with synthetic data (or hybrid augmentation) is an active area of interest.

CONCEPTUAL WORKFLOW / CASE EXAMPLE

Below is a conceptual workflow integrating open-source Python tools and/or an advanced solution, illustrating how synthetic data fits into a real trial development path:

PROTOCOL & METADATA INGESTION

- Extract key protocol elements: schedule, arms, endpoints, visits, inclusion criteria.
- Ingest CDISC metadata (e.g. define-xml, CRF specs, controlled terminology) as constraints.
- Configure target distributions (demographics, covariates, event rates).

SYNTHETIC DATA GENERATION

- Use open-source libraries and tools like – Python, Pandas
- Enforce constraints: visit ordering, date sequencing, code lists, derivations.
- Optionally inject anomalies/outliers to stress test QC logic.
- Output synthetic SDTM datasets plus supplemental ADaM-like derivative datasets.

VALIDATION & METRICS

- Compute similarity metrics (e.g. marginal KS test, correlation matrices, distribution overlaps).
- Assess utility: run a pilot analysis (e.g. summary table or model) and compare to expected benchmarks.
- Evaluate privacy: record risk scores, membership inference checks, or differential privacy metrics.

PROGRAMMING PIPELINE DEVELOPMENT

- Use synthetic data to build edit check logic, derivation macros, mapping modules, TFL skeletons.
- Iteratively refine, debug, and document.

DRY-RUN EXECUTION & QC

- Execute full pipeline on synthetic data: mapping → ADaM → TFL generation.
- Examine errors, performance, edge-case failures, logic gaps.
- Adjust code before real data arrives.

TRANSITION TO REAL DATA

- When real data begins arriving, execute code with minimal change.
- Differences should reflect only data-specific adjustments; bugs and architecture should already be stable.
- Retain synthetic-to-real comparators (e.g. track metrics, divergence).

ONGOING HYBRID USE

- Use synthetic data to test patches, new endpoints, protocol amendments.
- Use scenario synthetic cohorts to test regulatory sensitivity analyses.

Figure (suggested): A pipeline diagram showing “Protocol/Metadata → Synthetic Generator → Validation → Programming Pipeline → Dry Run → Real Data Adaptation”.

The solution must explicitly support features like outlier injection, controlled terminology compliance, and visit ordering, so it can serve as a production-grade synthetic engine in this workflow.

BENEFITS & REGULATORY IMPLICATIONS

BENEFITS & VALUE PROPOSITION

BENEFIT	DESCRIPTION
Time-to-deliverable acceleration	By front-loading development on synthetic data, teams reduce late-stage debugging and delays.
Cost efficiency	Fewer manual iterations, less rework, and fewer full-blown system crashes.
Quality & robustness	Exposure of edge cases and anomalies early improves resilience of programming pipelines.

BENEFIT	DESCRIPTION
Privacy-safe experimentation	Synthetic data provides a sandbox free from patient confidentiality constraints.
Bias stress testing	You can simulate underrepresented subpopulations or adverse scenarios to challenge downstream modules.
Continuous iteration	Synthetic datasets can be refreshed or varied to support protocol amendments, new endpoints, or substudies.

From a sponsor's or data scientist's lens, synthetic data shifts risk left: instead of discovering issues late (post–data-lock), you surface them early, reduce surprises, and shorten the critical path.

REGULATORY & COMPLIANCE CONSIDERATIONS

- **Not a substitute for final analysis data**: Regulators currently expect actual data for primary inference, safety analyses, and submission. Synthetic data is not yet a "drop-in" replacement.
- **Augmentation & hybrid use**: Synthetic data may inform simulations, sensitivity analyses, or external control arm augmentation, but with transparent disclosure and validation.
- **Submission transparency**: If synthetic data is used in development workflows, it should be documented—metadata, versioning, constraints, generation methods, and limitations.
- **Auditability**: Synthetic pipelines must be auditable (seed, randomization logs, parameters, code).
- **Regulatory precedent & guidance**: While guidance is nascent, some agencies are exploring AI-driven synthetic data use, especially in digital health or decentralized trials contexts (PHUSE working groups are exploring knowledge gaps).
- **Standards alignment**: Synthetic data workflows must respect CDISC (SDTM/ADaM) conventions, variable naming, controlled terminology, derivation rules, and data quality rules. Tools that generate CDISC-compliant synthetic datasets simplify this alignment.
- **Validation expectation**: Regulatory reviewers may expect justification of synthetic methods, comparison of synthetic vs real divergence, and error bounds.

To summarize, synthetic data is not a magic compliance pass—but when used responsibly, it becomes a powerful tool to de-risk development, accelerate delivery, and improve resilience.

FUTURE OUTLOOK

AI & GENERATIVE MODELING ADVANCES

As generative AI evolves, domain-specific models pretrained on large trial datasets will emerge. Models with built-in explainability, uncertainty quantification, and privacy constraints will boost trust and adoption.

REAL-WORLD DATA (RWD) FUSION & HYBRID APPROACHES

Future workflows may merge synthetic data with real-world data (e.g. registries, EHR) to generate "blended" datasets that maximize realism while preserving privacy. Synthetic sidecar arms or augmented controls may become more accepted in regulatory settings.

AUTOMATION & HUMAN-IN-THE-LOOP QC

Synthetic pipelines should increasingly be integrated into automated CI/CD (continuous integration / continuous deployment) frameworks, with human review gates for edge-case validation.
For example, in recent PHUSE submissions, automation of derive/TFL code with metadata-driven engines is advancing.

REGULATORY MODERNIZATION

As regulatory agencies gain exposure to synthetic methodologies, formal guidance and acceptance pathways may emerge—especially for synthetic data used in analytical validation, pre-submission simulations, or hybrid control design.

EDUCATION & CULTURAL ADOPTION

Wider adoption requires education of biostatisticians, programmers, and regulators—shifting mindsets from "data comes first" to "design-first, data-simulated, data-validated." Case studies and published success stories will drive momentum.

CONCLUSION

In the fast-paced environment of clinical development, waiting for final data to build, test, and validate programming pipelines is a liability. Synthetic data offers a pragmatic bridge: realistic, privacy-safe data that enables early development, rigorous testing, and robust programming before the first patient is enrolled. While synthetic data is not (yet) a substitute for real data, when used judiciously—within transparent, validated, and governed workflows—it can dramatically accelerate trial readiness, reduce rework, and improve pipeline stability.

The power of synthetic data lies not merely in generation, but in integration: combining domain logic, governance, validation, and human oversight. As generative AI, sharing standards, and hybrid approaches mature, synthetic data will move from a promising experiment to a foundational tool in modern clinical trials.

REFERENCES

1. **GenInvo.** *How Meaningful Synthetic Data Generation Tools Are Transforming AI Development.* https://geninvo.com/how-meaningful-synthetic-data-generation-tools-are-transforming-ai-development/
2. **GenInvo.** *Synthetic Data vs. Real Data.* https://geninvo.com/synthetic-data-vs-real-data/
3. **Liu S et al.** *Evaluation of Synthetic Tabular Data: Utility, Privacy, and Fidelity Trade-Offs.* arXiv preprint arXiv:2504.18544 (2025).
4. **Wang J et al.** *Synthetic Longitudinal Data Generation: State of the Art and Future Directions.* arXiv preprint arXiv:2309.12380 (2023).

RECOMMENDED READING

1. **GenInvo.** *How Synthetic Data Accelerates Drug Discovery in the Pharmaceutical Industry.* https://geninvo.com/how-synthetic-data-accelerates-drug-discovery-in-the-pharmaceutical-industry/
2. **GenInvo.** *Synthetic Patient Data in Clinical Trials – Why It's Important to Have Meaningful Synthetic Data.* https://geninvo.com/synthetic-patient-data-in-clinical-trials-why-its-important-to-have-meaningful-synthetic-data/

CONTACT INFORMATION (HEADER 1)

Your comments and questions are valued and encouraged.

Contact the author at:
Author Name: Abhijeet Arora
Company: GenInvo
Email: abhijeet.216@geninvo.com
Website: https://www.geninvo.com