A Pilot Assessment of LLM-Generated Synthetic Cohorts: A First Step Toward Robust Synthetic Control Arms

Manuel Cossio,[12] Anupama Vasudvan,[3] Deepa Jahadirdar,[3]
[1]Cytel, Inc., Geneva, Switzerland; [2]Universitat de Barcelona, Barcelona, Spain; [3]Cytel, Inc., Cambridge, MA, US.

Background

- External Control Arms (ECAs) are increasingly used in clinical research (1)
- But, there are challenges in accessing suitable data, ensuring comparability to trial populations, and maintaining patient privacy.
- High-quality synthetic cohorts preserve key statistical properties while protecting sensitive information remains a critical need. Large language models (LLMs) provide new opportunities for synthetic data generation.
- Despite this promise, concerns remain around fidelity, reproducibility, and transparency—particularly in regulatory contexts.

.

Objectives

- This study evaluates two methodologies based on large language models (LLM) for generating synthetic clinical trial datasets intended for use in External Control Arms (ECAs).
- It compares direct generative output against automated code execution for statistical fidelity and reproducibility.

Methods

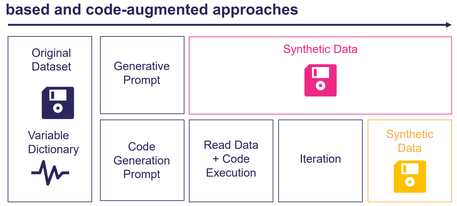
2.1 Data Source and Preprocessing

- As a reference dataset, this study used patient-level data from the Vitamin D and Omega-3 Trial (VITAL; NCT01169259), a large-scale randomized controlled trial sponsored by Brigham and Women's Hospital.
- A subset of the dataset was extracted and paired with a structured variable dictionary describing variable types (e.g., continuous, categorical), permissible ranges, and coding schemes (Table 1)

2.2 Direct LLM-Based Synthetic Data Generation

- The LLM was prompted to produce a fully synthetic dataset in Excel format. The model was provided with both the original dataset and the accompanying variable dictionary to inform the generation process.
- The prompt instructed the LLM to generate realistic patient-level records while maintaining consistency with the statistical structure, variable types, and constraints defined in the source data (Figure 1).

Figure 1. Framework for synthetic data generation using direct LLM-based and code-augmented approaches



Methods (cont.)

Table 1. Summary of Key Variables from VITAL Data Dictionary

Category	Variables (Examples)	Description
Study Identifiers	Subject_ID	Unique participant identifier
Treatment Assignment	vitdactive, fishoilactive	Randomization to Vitamin D and Omega-3 interventions (active vs placebo)
Demographics	sex, ageyr, raceth	Participant sex, age at randomization, and race/ethnicity categories
Clinical Baseline	bmi, currsmk, htnmed, diabetes	Baseline health characteristics including BMI, smoking, hypertension, and diabetes
Medication Use	cholmed, diabmed, statins, Aspirin	Baseline use of cholesterol, diabetes, statin, and aspirin medications
Outcomes & Events	majorcvd, malca, totmi, confdeath	Cardiovascular events, cancer outcomes, myocardial infarction, and mortality

2.3 Code-Augmented Synthetic Data Generation with Noise Injection

- The LLM was prompted to write code in Python to produce the synthetic cohort. This included resampling observations from the original dataset, followed by a total anonymization step using an algorithmic noise filter.
- Gaussian noise scaled to 5% of each continuous variable's original standard deviation was added. To preserve clinical plausibility, all perturbed values were clipped to their respective original ranges (Figure 1).

Results

- Both LLM-based approaches successfully generated synthetic cohorts of 100 patients derived from the VITAL trial dataset, enabling a direct comparison of generation strategy, statistical fidelity, and reproducibility.
- The direct generation approach was faster (Table 2), but the code-augmented introduced a structured and auditable workflow, allowing explicit control over each transformation step, including sampling, anonymization, and noise injection.
- Continuous variable generated from code-augmented synthetic cohort held patterns closer to the original data (Figure 2 and 3). These differences indicate that while direct generation captures general location, it may oversimplify underlying variability;
  - For age, the directly generated dataset exhibits a sharper, more concentrated peak around the mean, with reduced variance and a secondary elevation in the upper age range, suggesting less consistent modeling of the original distribution
  - For BMI, while the code-augmented approach reproduces the original structure with high fidelity, the directly generated dataset, shows a noticeably narrower and more peaked distribution, with diminished representation of higher BMI values and slight irregularities in the mid-range.

Results (cont.)

- The tails of both age and BMI variables reinforce the fidelity of the code-based approach as it preserves the gradual decline at higher values, which is important for edge cases or risk-relevant subpopulations.
- Explicit constraints on the code-augmented method drive its ability to preserve realistic distributions and values without excessive smoothing (Figure 2, Figure 3). In contrast, the direct LLM generation approach implicitly approximates distributions without formal constraints, leading to reduced variance and occasional distortion of distributional features.
- The reproducibility and full-traceability offered through the code-augmented approach may be particularly important in the context of external control arms, where regulatory acceptance depends on methodological clarity and the ability to validate data generation procedures.

Figure 2. Distribution of body mass index (BMI) in the original dataset compared with synthetically generated cohorts.

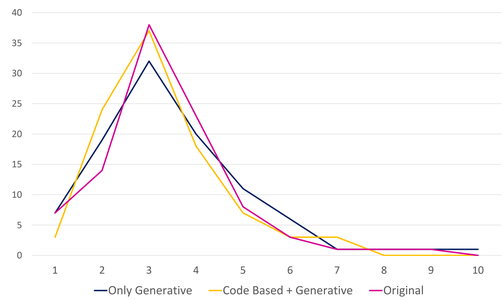


Figure 3. Distribution of age in the original dataset compared with synthetically generated cohorts.

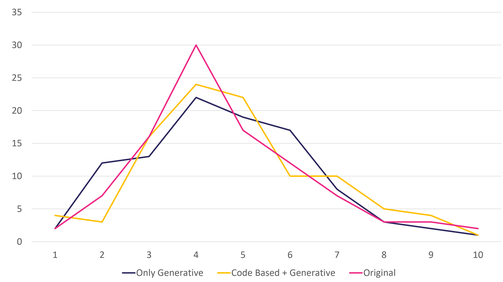


Table 2. Summary of Comparative Performance)

Dimension	Direct Generation	Code-Augmented Generation
Runtime	~23 seconds	~40 seconds
Iterations Required	1	12
Ease of Use	High	Moderate
Reproducibility	Low	High
Statistical Fidelity	Moderate	High
Transparency	Low	High
Privacy Control	Implicit	Explicit (noise parameterized)

Conclusions

Overall, both approaches demonstrate feasibility for generating synthetic clinical cohorts at small scale.

- While direct LLM generation offers rapid prototyping, code-based generation provides the transparency and granular statistical control essential for regulatory-grade external control arms.
- Calibrated Gaussian noise effectively balances data privacy with the preservation of population-level characteristics in trial-derived datasets.
- The visual alignment observed in age and BMI distributions (Figure 2, Figure 3), along with stable categorical proportions, supports the conclusion that structured, noise-based generation offers a more reliable pathway toward synthetic datasets suitable for downstream clinical and regulatory applications.
- Future work should systematically evaluate re-identification risk under adversarial attack models and compare noise-based anonymization against alternative privacy-preserving techniques such as differential privacy.

References

1 - Schmidli, Heinz, et al. "Beyond randomized clinical trials: use of external controls." Clinical Pharmacology & Therapeutics 107.4 (2020): 806-816.

2- Goyal, Mandeep, and Qusay H. Mahmoud. "An llm-based framework for synthetic data generation." 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2025.

Disclosures and acknowledgements

The study was investigator initiated. All authors are employees of Cytel, Inc. MC is also a researcher at Universitat de Barcelona.