Generative AI Secures Synthetic Datasets of NSCLC Trial Cohorts for RWD Analysis: A Privacy-Preserving Case Study

Manuel Cossio,[1,2] Ramiro Gilardino,[3]
[1]Cytel, Inc., Geneva, Switzerland; [2]Universitat de Barcelona, Barcelona, Spain; [3]Universidad de Buenos Aires, Buenos Aires, Argentina.

Background

- Health Economics and Outcomes Research (HEOR) is often limited by "data silos," where detailed patient-level data from clinical trials remains inaccessible due to privacy and proprietary constraints (1,2). This restricts researchers' ability to generalize findings or perform robust comparative analyses. As demand for Real-World Evidence (RWE) increases, methods are needed to bridge the gap between published summaries and the granular data required for health technology assessments.
- Generative AI offers a solution by creating synthetic datasets (3). Using Large Language Models (LLMs) to interpret trial characteristics and generate structured code, researchers can build "digital twins" of trial cohorts. These datasets retain key statistical properties—such as those of the CHRYSALIS-2 NSCLC cohort—without exposing patient data. This study evaluates a two-agent LLM pipeline that automates this process and assesses its readiness for clinical data modeling.

Objectives

- To assess whether a two-agent LLM pipeline can reliably extract and synthesize detailed clinical trial population parameters into privacy-preserving synthetic cohorts suitable for HEOR. The study addresses a persistent HEOR challenge: limited access to granular trial data due to patient confidentiality constraints, which restricts evidence generalization for value assessment and RWE analyses.

Methods

Framework Design and Initial Extraction

- The methodology employed a sequential two-agent LLM framework to transition from unstructured text to structured synthetic data. In Iteration 1, the first agent extracted core patient characteristics—median age, sex, race, and ECOG performance status. These outputs fed into a second agent, which generated Python code using NumPy and Pandas to simulate a dataset of N=105 patients. The code was evaluated across six dimensions: Fidelity & Correctness, Readability, Maintainability, Reproducibility, Documentation, and Efficiency.

Iteration 2: Enhanced Molecular Granularity

- The pipeline was refined in Iteration 2 through improved prompt engineering to capture complex mutational data. While Iteration 1 identified broad mutation groups such as Exon 18 G719X (56%) and Exon 21 L861X (26%), it lacked sufficient granularity. The second iteration addressed this by generating specific mutation counts and subtypes, splitting G719X into distributions of G719S, G719A, and G719C to better reflect clinical reality.

Table 1. Demographics and Clinical Characteristics (N = 105)

Characteristic	Value	Characteristic	Value
Total Sample Size	105	Brain Metastases (Yes)	37 (35.2%)
Median Age (Range)	64 (30–85)	ECOG PS 0	33 (31.4%)
Sex (Male)	53 (50.5%)	ECOG PS 1	72 (68.6%)
Sex (Female)	52 (49.5%)		
Race (Asian)	71 (67.6%)		
Race (White)	31 (29.5%)		
Race (Other)	3 (2.9%)		

Abbreviations: ECOG, Eastern Cooperative Oncology Group, PS, Performance Status.

Methods (cont.)

Table 2. Genomic Profile and Treatment History (N = 105)

Characteristic	Value	Characteristic	Value
EGFR Exon 18 (G719X)	56 (53.3%)	Treatment-Naïve	49 (46.7%)
EGFR Exon 21 (L861X)	26 (24.8%)	Previously Treated	56 (53.3%)
EGFR Exon 20 (S768X)	23 (21.9%)	Prior Afatinib (of treated)	34/56 (60.7%)
Compound Mutations (Yes)	29 (27.6%)		

Abbreviations: EGFR, Epidermal Growth Factor Receptor.

Reproducibility and Scalability Standards

- Reproducibility and scalability were treated as core requirements rather than optional enhancements within the framework. To guarantee consistent and verifiable results, strict computational standards were enforced throughout the analysis pipeline. In particular, the use of fixed random seeds (e.g., np.random.seed(42)) ensured that patient cohort generation and any stochastic processes could be exactly replicated across different runs and environments. This approach minimizes variability introduced by randomness and allows independent researchers to validate findings with confidence, strengthening the overall credibility of the study.
- In parallel, the methodology was designed with scalability in mind to accommodate substantially larger datasets beyond the scope of the initial analysis. By prioritizing vectorized operations over iterative procedures, the framework achieves greater computational efficiency and performance. While the current study focused on a relatively small cohort, these design choices ensure that the same workflow can be seamlessly extended to datasets exceeding 10,000 patients without significant modifications. As a result, the framework remains both practical for exploratory studies and robust enough for large-scale clinical or real-world data applications.

Results

Replication of Core Clinical Distributions

- The initial iteration showed strong performance in reproducing high-level demographic and clinical characteristics. Key variables such as ECOG performance status (31% PS 0, 69% PS 1) and baseline brain metastases (35% Yes) were accurately mirrored in the synthetic dataset. This indicates that the LLM pipeline is well-suited for capturing structured, aggregate-level information from published sources.

Code Quality and Structural Performance

- In Iteration 1, the generated code demonstrated strong structural clarity and organization, reflecting a well-disciplined approach to program synthesis. The LLM decomposed the workflow into logical steps—data initialization, variable assignment, and dataset assembly—each separated into clean, sequential blocks. Inline comments were used effectively to describe the purpose of each section, lowering the barrier for human interpretation and modification. This made the code not only readable but also easy to extend for adjacent use cases, such as adding new covariates or adjusting cohort size. From a scalability standpoint, the reliance on vectorized operations and standard Python libraries (NumPy, Pandas) ensured that the same structure could be applied to larger datasets with minimal modification.
- Despite these strengths, the "Fidelity & Correctness" rating remained at a medium level due to limitations in how comprehensively the clinical characteristics were captured. The mutation logic, in particular, was overly simplified and excluded "other atypical mutations," leading to an incomplete representation of the cohort's molecular profile. This gap highlights a key limitation of the first iteration: while structurally sound, the code did not fully encode the complexity of the source data, suggesting that strong software quality alone does not guarantee domain completeness.

Results (cont.)

Improved Modularity and Mutation Representation (Iteration 2)

- The second iteration significantly enhanced both flexibility and biological realism within the pipeline. By introducing non-exclusive mutation columns, the framework was able to more accurately capture the presence of compound mutations (28%), which were not adequately represented under the rigid, mutually exclusive structure used in the initial version. This shift allowed multiple mutation types to be recorded simultaneously, aligning the data structure more closely with real-world genomic complexity. As a result, the representation of patient-level mutation profiles became more nuanced and reflective of observed biological variation. In addition, the modular design of the updated pipeline enabled the inclusion of atypical EGFR mutations that had previously been excluded due to structural limitations. This improvement expanded the scope of the dataset without requiring major restructuring of the overall workflow. By accommodating a broader range of mutation types, the pipeline increased both the depth and practical usability of the dataset. The revised approach therefore supports more comprehensive downstream analysis while maintaining consistency in data organization.

Advances in Documentation and Transparency

- Iteration 2 introduced a more deliberate and structured approach to documentation, moving beyond basic inline comments to include explicit annotations of modeling decisions. The generated code clearly flagged where assumptions, imputations, or inferred distributions were applied due to missing or underspecified information in the source publication. In some cases, the LLM distinguished between directly extracted values and synthetically derived ones, improving traceability. This level of transparency is particularly important in HEOR contexts, where downstream users—such as health economists or regulatory reviewers—must understand not only what was generated, but how and why. As a result, the code became more auditable and better aligned with reproducible research standards. This highlights improvements in how processes and results are recorded and communicated. It emphasizes clearer, more accessible documentation to support transparency and understanding.

Hallucination Risk and Fidelity Trade-offs

- The shift toward finer-grained mutation modeling exposed a key limitation of LLM-driven synthesis: the tendency to prioritize domain plausibility over strict source adherence. In expanding the "Specific Other Atypical Mutation" category, the model introduced clinically recognized mutations (e.g., Exon 19 Insertions, Non-solitary Exon 20 Insertions) that were not reported in the CHRYSALIS-2 dataset. While these additions may improve face validity for a general NSCLC population, they deviate from the source-specific evidence base. This creates a subtle but important risk: synthetic datasets may appear more complete or realistic than the underlying data justifies. Consequently, fidelity remained rated as "Medium," underscoring the need for guardrails—such as stricter prompt constraints or post-generation validation—to balance completeness with accuracy. This addresses the balance between reducing hallucination risk and maintaining fidelity. It underscores the inherent trade-offs involved in optimizing for accuracy versus completeness.

Technical Performance and Reproducibility

- Both iterations demonstrated strong adherence to computational best practices. The consistent use of fixed random seeds ensured that all stochastic elements—such as age sampling or categorical assignments—were exactly reproducible across runs. This is critical for verification, peer review, and iterative model development. Additionally, the reliance on vectorized operations in NumPy and Pandas minimized computational overhead, allowing rapid generation even as dataset complexity increased. Memory usage remained efficient due to the use of structured arrays and list-based initialization patterns. Together, these features indicate that the pipeline is not only functionally correct but also technically robust and scalable for larger simulations. This focuses on the system's technical capabilities and consistency of results. It stresses the importance of reproducibility as a core aspect of reliable performance.

Complexity vs. Readability Trade-off

- The enhancements introduced in Iteration 2 came at the cost of increased code complexity. As the logic expanded to handle non-exclusive mutation classes, compound mutations, and multiple conditional distributions, the resulting code relied more heavily on dense array manipulations and layered transformations. While efficient, this style reduced immediate readability, particularly for users less familiar with vectorized programming paradigms. The simpler, more linear structure of Iteration 1 was easier to follow but less expressive. This trade-off highlights a common tension in advanced data engineering: as models become more realistic and flexible, they often require more sophisticated—and less transparent—implementations. Addressing this may require complementary strategies, such as modularization, helper functions, or augmented documentation, to preserve usability without sacrificing capability.

Table 3. Example of a Generated Synthetic Dataset Illustrating the Structure and Characteristics of the Simulated Data

ID	Age	Sex	Race	ECOG PS	Brain Metastasis at Baseline	Prior Treatment Status	Primary EGFR Mutation Type	Compound Mutations	Previous Afatinib
1	83	Female	Asian	1	No	Previously-Treated	Exon 18 G719X	No	Yes
2	67	Female	Asian	1	No	Treatment-Naive	Exon 18 G719X	No	N/A
3	66	Female	Asian	1	No	Previously-Treated	Exon 21 L861X	No	Yes
4	80	Male	White	1	Yes	Treatment-Naive	Exon 18 G719X	No	N/A
5	70	Female	White	1	No	Previously-Treated	Exon 21 L861X	No	Yes
6	58	Female	Asian	1	Yes	Treatment-Naive	Exon 21 L861X	No	N/A
7	78	Male	Asian	1	Yes	Previously-Treated	Exon 18 G719X	No	No
8	57	Male	Asian	1	Yes	Previously-Treated	Exon 21 L861X	No	Yes
9	58	Male	Asian	1	Yes	Treatment-Naive	Exon 20 S768X	No	N/A
10	56	Male	Asian	1	No	Treatment-Naive	Exon 20 S768X	Yes	N/A
11	55	Male	White	1	No	Previously-Treated	Exon 20 S768X	Yes	No
12	67	Female	Asian	0	No	Treatment-Naive	Exon 18 G719X	No	N/A
13	61	Female	White	1	No	Previously-Treated	Exon 20 S768X	Yes	Yes

Abbreviations: ECOG, Eastern Cooperative Oncology Group, PS, Performance Status, EGFR, Epidermal Growth Factor Receptor.

Conclusions

- This two-agent LLM architecture offers a scalable, reproducible approach to generating privacy-compliant synthetic clinical trial populations, with potential applications in early HEOR modeling, RWE generalization, and evidence synthesis. However, the emergence of fabricated clinical features when modeling complex variables represents a material risk for decision-grade use. Future research should incorporate automated validation and constraint-based verification layers to ensure synthetic data fidelity before integration into high-stakes HTA and pricing analyses

References

1. Ye, Jingjing, and Lei Nie. "Use of Real-World Data (RWD) and Real-World Evidence (RWE)." Innovative Designs and Analyses for Small Population Clinical Trials: Development Strategies and Operational Engagement for Pediatric and Rare Diseases (2024): 351-362.
2. Grimberg, Frank, et al. "The real-world data challenges radar: a review on the challenges and risks regarding the use of real-world data." Digital Biomarkers 5.2 (2021): 148-157.
3. Long, Lin, et al. "On LLMs-driven synthetic data generation, curation, and evaluation: A survey." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

Disclosures and acknowledgements

The study was investigator initiated. MC is an employee of Cytel, Inc and a researcher of Universitat de Barcelona. RG is a professor at Universidad de Buenos Aires.