

DAILY ENGLISH READING

11 August 2026 · 30-minute active reading pack for Davin

TODAY'S QUESTION

How do you generate synthetic clinical-trial data without generating invalid clinical logic?

WHY SELECTED

This reading connects generative AI directly to CDISC testing. LLM4ODM generates synthetic study records from formal metadata and uses a separate deterministic validation layer. CDISC's current AI challenge extends the same idea to synthetic SDTM and ADaM test data with traceable linkage.

Difficulty: C1 · synthetic data + CDISC + validation + SP testing

READING ORDER · 30 MINUTES

Min	Task	Focus
0-3	Preview	Distinguish realistic data from valid test data.
3-14	LLM4ODM	Identify inputs, validation rules and reported results.
14-20	CDISC	Read synthetic SDTM/ADaM use case and judging criteria.
20-25	synadam	Review supported ADaM structures and relationships.
25-30	Output	Design a synthetic test dataset for one SP workflow.

SOURCES · OPEN ACCESS

MAIN · 21 May 2026

LLM4ODM: Synthetic Clinical Study Data Generation for CDISC ODM

[PubMed / full-text link](#)

SUPPORT 1 · CDISC 2026 AI Innovation Challenge

AI-Enabled Synthetic Data Generation for Automation Testing

[Official CDISC page](#)

SUPPORT 2 · CRAN · 22 Jul 2026

synadam: Generate Synthetic ADaM Datasets

[Official CRAN package](#)

BRIEF BACKGROUND

Clinical systems need test data for development, validation and regression testing, but real patient data are often difficult to use because of privacy and governance restrictions.

LLM4ODM converts CDISC ODM metadata into structured prompts. The model proposes synthetic subject records; a separate validation layer checks schema, data types, branching logic and temporal coherence.

Across ten ODM datasets, the authors report 100% schema adherence after validation and more than 78% reduction in manual effort, together with stronger clinical plausibility than rule-based generation under expert review.

CDISC's 2026 AI Innovation Challenge explicitly asks for synthetic SDTM and/or ADaM data with traceable linkage to source inputs and metadata. Standards integration and traceability receive the highest judging weight.

The synadam R package offers a complementary workflow that synthesizes ADSL, BDS, OCCDS and TTE datasets while preserving selected structures, ranges and relationships.

KEY VOCABULARY

Term	中文	Meaning / use
synthetic data	合成数据	Artificially generated data designed to resemble real data without representing real patients.
schema adherence	模式/结构遵循性	Whether generated records conform exactly to the required data structure.
clinical plausibility	临床合理性	Whether combinations and trajectories make medical sense.
branching logic	分支逻辑	Rules that determine which fields or visits should appear based on prior values.
temporal coherence	时间一致性	Whether dates and event order follow a logically possible sequence.
metadata-driven	元数据驱动的	Generated or controlled from formal metadata rather than free-form instructions alone.
traceable linkage	可追溯关联	A documented connection from generated values back to source rules or metadata.
constraint system	约束系统	A set of explicit rules that generated data must satisfy.
subject-level dataset	受试者级数据集	A dataset with one record per subject, such as ADSL.
longitudinal dataset	纵向数据集	Repeated observations for a subject across visits or time.
occurrence data	发生型数据	Event-like records such as adverse events or concomitant medications.
time-to-event	事件时间数据	Analysis data built around an event time and censoring status.
test fixture	测试夹具/测试数据样本	Controlled data prepared to test specific software or derivation behavior.
edge case	边界情况	An unusual but valid case that can expose weaknesses in logic.
data utility	数据效用	How useful synthetic data are for testing, modeling or analysis.

USEFUL PHRASES

1. generate data from formal metadata

The system generates data from formal metadata rather than an unconstrained prompt.

2. validate the output against explicit rules

Every generated record is validated against explicit rules.

3. preserve clinically meaningful relationships

Synthetic data should preserve clinically meaningful relationships.

4. separate generation from validation

The workflow separates generation from validation.

5. inject targeted edge cases

The test set should inject targeted edge cases.

6. retain traceability to source metadata

Each generated field retains traceability to source metadata.

7. test both happy paths and failure paths

Validation should test both happy paths and failure paths.

8. avoid confusing realism with correctness

Teams should avoid confusing realism with correctness.

9. use synthetic data as a controlled test asset

Synthetic data can be used as a controlled test asset.

10. regenerate reproducibly from versioned inputs

The dataset should be reproducible from versioned inputs.

COMPREHENSION QUESTIONS

1. Why is clinical plausibility not the same as schema validity?
2. What role does metadata play in LLM4ODM?
3. Why should generation and validation be separate stages?
4. What does traceable linkage mean in synthetic SDTM/ADaM generation?
5. How is synadam different from metadata-first generation?

RETELLING PROMPTS

30 seconds: Problem -> approach -> result.

45 seconds: Metadata -> generation -> deterministic validation -> synthetic records -> testing.

60 seconds: Apply the design to a SAS macro, ADaM derivation or TFL program.

5-MINUTE SPEAKING / OUTPUT TASK

Your role: Create a reusable synthetic test dataset for an SP team.

Minute 1: Choose ADSL, ADLB, ADAE, ADTTE or one TFL family.

Minutes 2-3: Define normal, missing, boundary, contradictory and rare edge cases.

Minute 4: List the deterministic rules that validate the generated data.

Minute 5: Explain acceptance criteria for regression-test use.

ONE SENTENCE TO KEEP

Synthetic clinical-trial data are valuable not because they look realistic, but because they are reproducible, traceable and intentionally constructed to test the rules that matter.