

DAILY ENGLISH READING

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

TODAY'S QUESTION

When can synthetic clinical-trial data support valid statistical inference?

WHY SELECTED

A new open-access Scientific Reports study stress-tests synthetic data across 115 phase 3 oncology trials. The key lesson for SP automation is that structural realism does not guarantee valid confidence intervals, p-values or treatment-effect conclusions.

Difficulty: C1 · synthetic data + oncology + statistical inference + CDISC testing

READING ORDER · 30 MINUTES

Min	Task	Focus
0-3	Preview	Realistic vs inferentially valid synthetic data.
3-15	Main	False positives, SE correction and power.
15-21	CDISC	Synthetic SDTM/ADaM for automation testing.
21-25	Retrieve	Method, intended use, metric, failure mode.
25-30	Output	Define fit-for-purpose acceptance criteria.

SOURCES · OPEN ACCESS

Main: Challenges of analyzing synthetic tabular data generated from 115 phase 3 oncology trials · Scientific Reports · 23 Jul 2026

[O a7ion and po- 50.4 cm 1ItaF0 1 6 66.2 cRnspjOkearp o cm 0 1g0 0 0 rg BT 1t oncology trials · Scientific Reports · 23 Jul 2026FTL /F2p*](#)

KEY VOCABULARY

Term	中文	Meaning / use
inferential utility	推断效用	Whether synthetic data can support statistically valid conclusions, not merely realistic-looking records.
treatment effect estimate	治疗效应估计值	A numerical estimate of the difference between treatment groups.
false positive	假阳性	A statistically significant result that appears when no corresponding effect should be concluded.
standard error	标准误	A measure of uncertainty around an estimated effect.
uncertainty propagation	不确定性传播	Carrying uncertainty from data generation into downstream statistical inference.
parametric model	参数模型	A model defined by a finite set of statistical parameters and distributional assumptions.
deep generative model	深度生成模型	A neural generative model such as a GAN or related architecture used to synthesize data.
power loss	统计功效损失	Reduced probability of detecting a true effect.
synthetic sample size	合成样本量	The number of generated observations used in a synthetic analysis.
naive analysis	朴素分析 / 未校正分析	An analysis that ignores extra uncertainty introduced by synthetic-data generation.
replicate	复现	To reproduce an original estimate or analysis result.
randomized oncology trial	随机肿瘤试验	An oncology trial assigning participants to treatment groups by randomization.
statistical validity	统计有效性	The extent to which inference follows appropriate statistical assumptions and error control.
privacy-sensitive context	隐私敏感场景	A setting where access to real patient-level data is restricted.
fit for purpose	适合预定用途	Appropriate for the specific task for which data or a method will be used.

USEFUL PHRASES

1. **preserve the original treatment-effect estimate** — *The synthetic analysis should preserve the original treatment-effect estimate.*
2. **account for uncertainty introduced by data generation** — *The method must account for uncertainty introduced by data generation.*
3. **inflate the false-positive rate** — *A naive workflow can inflate the false-positive rate.*
4. **correct the standard errors before inference** — *The authors correct the standard errors before inference.*
5. **trade statistical power for better error control** — *The method may trade statistical power for better error control.*
6. **increase the synthetic sample size** — *One way to recover power is to increase the synthetic sample size.*
7. **distinguish data realism from inferential validity** — *Teams must distinguish data realism from inferential validity.*
8. **validate each intended downstream use separately** — *Synthetic data should be validated for each intended downstream use separately.*
9. **treat generation and analysis as one statistical system** — *Generation and analysis should be treated as one statistical system.*
10. **define acceptance criteria before production use** — *Teams should define acceptance criteria before production use.*

COMPREHENSION QUESTIONS

1. Why can realistic synthetic data still produce invalid statistical inference?
2. What happens when generation uncertainty is ignored?
3. Why does standard-error correction improve error control but reduce power?
4. Why are parametric generators currently more promising for inferential utility?
5. How should validation differ for ADaM regression testing versus treatment-effect inference?

RETELLING PROMPTS

30 sec: Study scale → failure mode → conclusion.

45 sec: Generation → underestimated uncertainty → false positives → correction → power trade-off.

60 sec: Structural validity → clinical/logical validity → inferential validity.

5-MINUTE SPEAKING / OUTPUT TASK

Minute 1: Choose unit testing, macro regression, ADTTE logic, TFL testing, training or inference.

Minutes 2-3: Define structural, logical and statistical acceptance criteria.

Minute 4: Separate acceptable and unacceptable failure modes.

Minute 5: Explain what extra evidence is required for inferential use.

ONE SENTENCE TO KEEP

Synthetic clinical-trial data are fit for purpose only when the validation target matches the downstream decision: structural realism is not the same as inferential validity.