DAILY ENGLISH READING

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

TODAY'S QUESTION

Can an AI statistics agent reason in the cloud while patient-level data stay local?

WHY SELECTED

This fresh clinical-statistics preprint separates external AI reasoning from local row-level computation. The architecture is directly transferable to CRO/SP automation because data access, code execution, validation and human judgment can live behind different trust boundaries.

Difficulty: C1 · agentic AI + local execution + privacy + clinical statistics

READING ORDER · 30 MINUTES

Min	Task	Focus
0-3	Preview	What leaves the local environment?
3-14	Main	Architecture, local execution and reproduced analysis.
14-20	FDA/EMA	Context of use, governance and lifecycle.
20-25	ClinAgent	Compare domain skills and bounded tools.
25-30	Output	Design one SP data boundary.

SOURCES · OPEN ACCESS

Main · medRxiv · 30 Jul 2026
A Privacy-Preserving Zero-Code Conversational Statistical Analysis System for Clinical Research Using Agentic AI and Local R Execution
Open preprint

Support · FDA / EMA · Jan 2026
Guiding Principles of Good AI Practice in Drug Development
Official principles

Support · Biology Methods and Protocols
ClinAgent: AI-assisted methodology for clinical trial data processing and statistical programming
Open article

BRIEF BACKGROUND

The system separates an external language model from local row-level statistical computation. The authors report that the external model receives no row-level patient records.

The workflow includes schema perception, interactive cleaning, requirement refinement, retrieval-grounded statistical guidance, generated R code and controlled local execution.

Functional validation reproduced Kaplan-Meier analysis, Cox regression, diagnostics, ROC analysis and publication-ready outputs from a published prognostic study.

Keeping raw rows local is useful but not sufficient: prompts, metadata, logs and output artifacts also require deliberate privacy and governance controls.

For SP work, a practical pattern is: approved metadata -> LLM reasoning -> local SAS/R execution -> deterministic validation -> human approval -> auditable output.

KEY VOCABULARY

Term	中文	Meaning / use
privacy-preserving architecture	隐私保护架构	A system design that limits exposure of sensitive data while still enabling useful computation.
local execution	本地执行	Running code and processing row-level data inside a controlled local environment.
row-level data	行级数据	Individual patient or observation records rather than aggregated summaries.
schema perception	数据结构感知	Identifying variables, types, labels, missingness and other structural properties of a dataset.
requirements refinement	需求细化	Clarifying an analytical request before code is generated or executed.
controlled command-line interface	受控命令行接口	A restricted execution interface that exposes only approved operations.
retrieval-augmented generation	检索增强生成	Grounding an LLM with selected external knowledge or templates before it responds.
multivariable Cox model	多变量 Cox 模型	A survival model estimating hazard relationships while adjusting for multiple covariates.
Schoenfeld residual	Schoenfeld 残差	A diagnostic quantity commonly used to assess proportional-hazards assumptions.
publication-ready output	可直接用于发表的输出	A table or figure formatted to a standard suitable for formal reporting.
human-supervised	人工监督的	A workflow in which humans retain control over judgment-dependent decisions.
data boundary	数据边界	The defined point beyond which sensitive data are not allowed to move.
execution boundary	执行边界	The defined separation between reasoning, code generation and actual computation.
audit trail	审计追踪	A record of actions, inputs, code, outputs and approvals that allows later reconstruction.
context of use	使用情境	A precise definition of why, where and for what purpose an AI system is used.

USEFUL PHRASES

1. **keep raw patient data inside the local environment** - *The architecture keeps raw patient data inside the local environment.*

2. **separate remote reasoning from local execution** - *The system separates remote reasoning from local execution.*

3. **send metadata rather than patient-level records** - *The agent can send metadata rather than patient-level records.*

4. **ground statistical choices in curated guidance** - *Statistical choices are grounded in curated guidance.*

5. **require human confirmation for judgment-dependent decisions** - *The workflow requires human confirmation for judgment-dependent decisions.*

6. **execute generated code through a controlled interface** - *Generated code is executed through a controlled interface.*

7. **reproduce a published analytical workflow** - *The framework was tested by reproducing a published analytical workflow.*

8. **treat privacy as an architectural constraint** - *Privacy should be treated as an architectural constraint.*

9. **log every analytical decision and execution step** - *The system should log every analytical decision and execution step.*

10. **define the model's context of use before deployment** - *Teams should define the model's context of use before deployment.*

COMPREHENSION QUESTIONS

1. Why separate external reasoning from local execution?

2. What information can still leak sensitive content even when raw rows stay local?

3. Why should judgment-dependent statistical choices require human confirmation?

4. How do FDA/EMA principles extend beyond privacy architecture?

5. How could this design transfer to SAS/CDISC statistical programming?

RETELLING PROMPTS

30 sec: Privacy problem -> split architecture -> reproduced analysis.

45 sec: Requirement -> external reasoning -> local execution -> validation -> approval.

60 sec: Explain why no raw rows in the LLM is necessary but not sufficient.

5-MINUTE SPEAKING / OUTPUT TASK

Minute 1: Choose one SP AI use case.
Minutes 2-3: Separate model-safe, local-only and human-approval information.
Minute 4: Add allowlisted tools, local execution, logging and deterministic QC.
Minute 5: State the minimum information the model truly needs.

ONE SENTENCE TO KEEP

A trustworthy clinical AI workflow should expose the model to the minimum information needed for reasoning, keep sensitive computation inside a controlled execution boundary, and preserve human accountability for judgment-dependent decisions.