DAILY ENGLISH READING

14 September 2026 - 30-minute active reading pack

TODAY'S QUESTION

When free-text EHR data becomes analysis data, what must an SP preserve?

Difficulty: C1 | LLM extraction | EHR / RWD | FHIR-to-CDISC | provenance

WHY SELECTED

Nature Medicine shows how LLMs can extract computable clinical entities from free text, validate them with physician review, combine them with structured EHR data and analyze longitudinal journeys. For SP work, the extraction itself is a derivation whose source evidence, timing, terminology and provenance must survive downstream.

READING ORDER

0-3 min Preview: text -> structured entity -> knowledge graph.

3-14 min Main: adjudication, ontologies, trajectories and note-only outcomes.

14-20 min CDISC RWD: event-based EHR data vs protocol-based research data.

20-25 min FDA: quality considerations for EHR and claims RWD.

25-30 min Output: build one traceable free-text-to-analysis derivation.

CORE PIPELINE

source note -> extracted concept -> validation / adjudication -> terminology

-> temporal anchor -> research representation -> analysis variable -> result

Do not erase the source sentence when the concept becomes structured.

Do not assume note date equals event date.

Treat an AI-derived clinical fact as derived data, not raw source.

SP CONNECTION

EHR data are often event-based; research datasets are usually protocol-based.

FHIR-to-CDISC mappings help bridge healthcare and research semantics.

Missingness, irregular collection and heterogeneous dictionaries remain important.

ADaM/TFL review should be able to trace a derived result back to extraction evidence.

KEY VOCABULARY

unstructured clinical text	非结构化临床文本	narrative EHR notes and descriptions
computable clinical data	可计算临床数据	structured information software can analyze
physician adjudication	医生裁定	expert review of extraction correctness
blinded reference standard	盲态参考标准	gold standard independent of prediction
inter-reviewer agreement	审阅者间一致性	agreement between independent reviewers
medical ontology	医学本体	standardized clinical concepts and relations
knowledge graph	知识图谱	entities connected by queryable relationships
phenotype extraction	表型提取	identify clinical characteristics from records
longitudinal trajectory	纵向轨迹	patient measurements and outcomes over time
temporal anchoring	时间锚定	assign event to correct clinical time
time-to-event outcome	事件时间结局	time from origin to event or censoring
data heterogeneity	数据异质性	variation across sites systems and collection
semantic mapping	语义映射	preserve meaning across data models
source provenance	来源谱系	where an extracted or derived value came from
analysis-ready representation	分析就绪表示	structured validated data ready for analysis

USEFUL PHRASES

1. turn narrative evidence into computable variables

2. preserve the source sentence behind every extracted concept

3. temporal alignment is part of the derivation

4. structured and unstructured data should be reconciled rather than treated as separate worlds

5. a clinically plausible extraction still requires validation

6. the ontology provides shared meaning across heterogeneous records

7. the knowledge graph enables patient-level and population-level queries

8. research data are protocol-based while EHR data are event-based

9. missingness in routine care has a different mechanism from planned trial missingness

10. an analysis-ready variable should carry its provenance forward

COMPREHENSION QUESTIONS

1. Why can notes contain information that structured EHR fields miss?

2. Why is physician adjudication important for LLM-created variables?

3. What do ontologies and knowledge graphs add after extraction?

4. Why is temporal anchoring difficult in real-world data?

5. What provenance should survive if the concept reaches ADaM or a TFL?

RETELLING PROMPTS

30 sec: note -> extracted concept -> structured data -> longitudinal analysis.

45 sec: event-based healthcare data vs protocol-based research data.

60 sec: explain why an AI-extracted EHR variable is a derivation.

5-MINUTE SPEAKING / OUTPUT TASK

Minute 1: choose one free-text clinical statement.

Minutes 2-3: define entities, timing, terminology, target structure and analysis use.

Minute 4: separate deterministic QC from human clinical judgment.

Minute 5: explain what provenance must survive into ADaM / TFL review.

ONE SENTENCE TO KEEP

When an LLM converts clinical narrative into analysis data, the extraction itself becomes a derivation, so source evidence, timing, terminology, validation and provenance must travel with the value downstream.