The Chart is Not the Patient: Benchmarks Underrate the Evidence Problem in Clinical AI

David M. Vidmar[1], Nicole Hakim[1], Greg Lee[1], John Pfeifer[1], Logan Brigman[1], Christopher M. Haggerty[1], Brandon K. Fornwalt[1]

[1] Tempus AI, Inc., Chicago, IL

INTRODUCTION

Current clinical agent benchmarks largely focus on **clinical uncertainty** i.e., the ability to reason **from** the record

Due to the messy nature of the EHR, evaluations **about** the record i.e., **clerical uncertainty** are needed to improve estimates of real-world performance

SYSTEMIC ORIGINS

Clerical uncertainty is inherent to healthcare data

- **Multiple motives**
  billing, admin, clinical
- **Habitual stagnation**
  copy-forward problem list
- **Fractured information**
  Department contradictions

Example of Multiple Motives
Only 16% of patients coded for pulmonary hypertension were confirmed by chart review to have a clinical diagnosis[1]

CURRENT BENCHMARKS

I. Static exams: clean vignettes, taken as fact
MedQA, MultiMedQA

II. Patient dialogue: facts may be missing, not wrong
CRAFT-MD, MediQ, HealthBench

III. EHR agents: real records, but clerical skill is never isolated
MedAgentBench, MedHELM

DETECTION IS NOT ENOUGH

Unlike error-detection tasks, evaluations must grade the complete response to a defect, not just its detection

1. Gillmeyer KR, et al. Chest. 2021;159(5):1986-1994.

A DEFECT-RESPONSE GRID FOR CLERICAL UNCERTAINTY

	Proceed Ignore as unimportant	**Reconcile** Use other data to resolve	**Hedge** Act with lower confidence	**Clarify** Escalate to another source
Completeness Is the evidence present? e.g., missing documentation	Out-of-scope test missing → Not needed	No echo; measurement cited in a note → Use measurement	Risk score input missing → Label as estimate	No outside records → Request the records
Correctness Is the value accurate? e.g., unit errors	Lab without unit in expected range → Assume unit	Weight in lb entered as kg → Fix using trend	Doubtful value near a cutoff → Choose the safer option	Implausible drug dose → Confirm with prescriber
Currency Does it still hold? e.g., copy-forward notes	Social history copied forward → Irrelevant to task	Problem list “active”; new notes show resolved → Trust notes	Latest vitals months old → Report with their date	Old med still listed (years) → Ask patient
Credibility Any reason to doubt this? e.g., notes contradict codes	Rule-out code, out of scope → Not needed	Test result refutes a code → Trust the test	Diagnosis with no workup → Treat as unconfirmed	Notes conflict on care goals → Ask care team

Each cell shows an example whose correct response, set by clinicians, is that column.

MORE TRUSTING ⟷ **MORE SKEPTICAL**

P R H C

Always proceeds
Right in 4 of 16 cells
Takes the chart at face value

Calibrated
Right in all 16 cells
What competence requires

Always clarifies
Right in 4 of 16 cells
Over-asks, inviting alert fatigue

Each defect type hits a known LLM weakness: under-flagging absence, weak plausibility checks, weak temporal reasoning, over-weighting repeated claims

IMPLEMENTATION

Perturb real longitudinal EHRs with clinician-defined defects then grade the response

EXAMPLE TEST
Perturb: flip a resolved problem-list entry to "active" (Currency)
Observe: does the agent proceed, reconcile against notes, or ask?
Grade: its final action vs. the clinician-defined expectation

Metric: non-inferiority to clinicians, who are imperfect here too

CONCLUSION

For better real-world alignment, clinical agents should be tested on **clinical reasoning** from data with **clerical uncertainty**

ACKNOWLEDGMENTS
We thank Dana DeSantis from the Tempus Science Communications team for poster development.

Correspondence:
david.vidmar@tempus.com
greg.lee@tempus.com