

HealthRadar Yardstick 2.0

Effect of structured clinical context on material failures in large-language-model laboratory interpretations

Edward A. Rose, M.D., M.S.A. and Steven E. Rose, Ph.D.

Public manuscript • Version 1.0 • 5 October 2026

HealthRadar research program • <https://healthradarmvp.com/>

Abstract

Background: Patient-facing explanations of laboratory results can be fluent and factually plausible while still misstating what matters most clinically, including the appropriate action window. HealthRadar was developed as a governed clinical-context layer intended to preserve clinical priorities, uncertainty, treatment boundaries, and routing information before a generative model produces a patient-facing interpretation.

Objective: To compare material clinical failure in reports generated from the same synthetic laboratory cases with a common raw prompt versus the same prompt supplemented by structured HealthRadar clinical context.

Methods: Fifteen prospectively defined synthetic laboratory cases were evaluated across three frontier-model families, three repetitions, and two paired study arms (Raw and Guided), yielding 270 reports. Case facts, physician reference standards, common generation instructions, augmentation rules, and the primary analysis were frozen before the final comparison. Material clinical failure was defined by the locked overall classification. The primary effect was the case-balanced Guided-minus-Raw difference. Uncertainty was estimated with 50,000 paired case-cluster bootstrap replicates. Clinical ratings were AI-assisted and physician-confirmed by one physician associated with HealthRadar.

Results: Material clinical failure occurred in 18/135 Raw reports (13.33%) and 2/135 Guided reports (1.48%). The case-balanced Guided-minus-Raw difference was -11.85 percentage points; the 95% paired case-cluster percentile interval was -25.19 to 0.00 percentage points. The favorable difference was concentrated in three safety-sensitive cases: low sodium, severe neutropenia, and severe hypertriglyceridemia. Twelve of fifteen case-level effects were zero. Across both arms, 141 reports were classified Excellent, 109 Acceptable, 20 Materially deficient, and none Potentially unsafe. Eighteen of the 20 material failures concerned failure to preserve the prespecified timing or disposition boundary; two concerned categorical exclusion of iron deficiency or iron contribution in a CKD/anemia case.

Conclusions: In this internal synthetic benchmark, structured HealthRadar context was associated with a large descriptive reduction in material clinical failure. The confidence interval reaches zero, the case set was small and purposive, and scoring was not independently performed by multiple external clinicians. The study therefore does not establish real-world effectiveness, patient benefit, or general superiority of HealthRadar. It provides a testable signal for external replication using independently constructed cases and independent clinical scoring.

Introduction

Large language models can generate coherent explanations of laboratory results, but the clinically important task is not simply to restate abnormalities. A useful report must identify the principal finding, calibrate urgency,

preserve uncertainty, avoid unsupported diagnoses or treatment directives, and communicate a proportionate next step.

HealthRadar Yardstick 2.0 was designed around a narrow question: when the underlying patient and laboratory facts are held constant, does adding a structured, physician-governed clinical context layer reduce material clinical failures in downstream generative-model reports? The study was not designed to estimate population error prevalence, demonstrate patient benefit, or validate a deployed medical product.

The paired design also addresses a common evaluation problem. Repeated model outputs from the same case are not independent patients. The primary analysis therefore treats the clinical case, rather than each report, as the key unit for uncertainty estimation.

Methods

Study design

Yardstick 2.0 used 15 synthetic cases (Cases 21-35), three model families, three stateless repetitions per model, and two paired arms. Each case-model-repetition combination was generated once in the Raw arm and once in the Guided arm, producing 135 reports per arm and 270 reports overall.

The Raw arm received the frozen case facts and common generation instructions. The Guided arm received the same materials plus a structured HealthRadar augmentation containing the case scope, principal and supporting findings, routing information, safety drivers, uncertainty and capability boundaries. The comparison therefore estimates the effect of the full additional-context package; it does not isolate the contribution of any single rule.

Clinical cases and reference standards

The case set intentionally spanned routine or near-normal results, mild nonurgent abnormalities, risk-marker patterns, anemia and iron deficiency, inflammatory findings, concordant and discordant glycemic patterns, electrolyte abnormalities, advanced chronic kidney disease, severe neutropenia, and severe hypertriglyceridemia. It was a deliberately enriched challenge set rather than a prevalence sample.

Each case had a physician reference specifying the principal interpretation, acceptable supporting content, required disposition or follow-up, uncertainty and treatment boundaries, safety-netting, and conditions that would constitute material deficiency or a potentially unsafe report.

Scoring and primary endpoint

The primary endpoint was material clinical failure, operationalized from the frozen overall report classification. The overall categories were Excellent, Acceptable, Materially deficient, and Potentially unsafe. Material clinical failure comprised the latter two categories for the primary analysis.

The review process was AI-assisted and physician-confirmed. The assistant proposed report-level categories, domain scores, and structured judgments; Edward A. Rose, M.D., M.S.A., the study's clinical owner and a HealthRadar co-founder, reviewed and confirmed the clinical scoring. This was not independent, unaided physician scoring of all 270 reports, and interphysician reliability was not measured.

Statistical analysis

For each case, failure proportions were first calculated within the model-by-arm triplicates and then compared between Guided and Raw while preserving the paired case structure. The primary estimate was the case-balanced Guided-minus-Raw difference in material-failure proportion, expressed in percentage points.

A paired case-cluster percentile bootstrap resampled whole cases 50,000 times using the frozen analysis seed 20260930. All model families, arms, and repetitions for a selected case remained together within each resample. The reported 95% interval is the 2.5th to 97.5th percentile of that distribution. It is not a Bayesian posterior interval and the bootstrap draws do not create additional independent clinical cases.

Computational verification

A separate read-only computational audit reproduced the score/crosswalk join, the primary endpoint, the case-balanced effect estimate, and all 50,000 saved bootstrap values using an independently implemented calculation. This verifies the arithmetic and data linkage; it is not an independent clinical validation because the verifying assistant also participated in study support.

Results

Primary result

Arm	Failures	Reports	Failure proportion	Case-balanced contrast
Raw	18	135	13.33%	Reference
Guided	2	135	1.48%	-11.85 percentage points

The 95% paired case-cluster percentile interval for the Guided-minus-Raw difference was -25.19 to 0.00 percentage points.

Distribution across cases

Case pattern	Raw failures / 9	Guided failures / 9	Guided - Raw
Cases 21-31 (each)	0	0	0.00 pp
Case 32: low sodium	6	0	-66.67 pp
Case 33: CKD/anemia	1	1	0.00 pp
Case 34: severe neutropenia	4	0	-44.44 pp
Case 35: severe triglycerides	7	1	-66.67 pp

Only three of 15 case-level effects were nonzero, and all three favored the Guided arm. Twelve cases had a zero Guided-minus-Raw effect, including the CKD/anemia case in which each arm had one failure.

Because 12/15 case effects were zero, a bootstrap resample containing only zero-effect cases had nontrivial probability. In the frozen 50,000-replicate run, 1,811 resamples produced an effect of exactly zero, which is why the upper limit of the 95% interval is exactly 0.00 rather than a rounded negative value.

Overall classifications and failure mechanisms

Across all 270 reports, 141 were classified Excellent, 109 Acceptable, 20 Materially deficient, and 0 Potentially unsafe. Excellent classifications were nearly identical by arm (71 Raw, 70 Guided), arguing against a simple explanation in which guidance merely made prose appear more polished.

Eighteen material failures concerned action timing or disposition in Cases 32, 34, and 35. Two concerned categorical exclusion of iron deficiency or iron contribution in Case 33. The timing endpoint reflects conformity with prospectively selected clinical action windows; the benchmark did not observe whether any patient delayed care or experienced harm.

Model-family descriptive results

Model family	Raw failures / 45	Guided failures / 45
OpenAI	1	0
Anthropic	8	1
Google	9	1

These family-specific counts are descriptive. Yardstick 2.0 was not powered to establish superiority or equivalence among individual model families.

Discussion

The central Yardstick 2.0 observation is not that HealthRadar made every report better. Most cases had no material failure in either arm. The observed benefit was concentrated where the reference standard required preservation of a time-sensitive clinical boundary. That pattern is consistent with the intended function of a governed context layer: to carry prioritization and action constraints into the generation step.

At the same time, the concentration of benefit limits generalization. With only 15 distinct cases, three cases account for the entire favorable case-level contrast. The confidence interval reaches zero, so the study should not be described as establishing statistically significant superiority under the reported two-sided interval.

The equal Excellent totals also matter. A benchmark intended to measure clinically consequential failures should not confuse more polished language with safer clinical interpretation. In Yardstick 2.0, the primary signal came from material-failure classifications rather than a shift toward more Excellent prose.

Limitations

- Synthetic, purposively constructed cases cannot estimate real-world error prevalence or patient benefit.
- Only 15 distinct cases contributed independent clinical variation to the primary case-level analysis.
- The favorable effect was concentrated in three safety-sensitive cases.
- Clinical scoring was AI-assisted and physician-confirmed by one HealthRadar-associated physician; independent external clinician scoring and interrater reliability were not measured.
- The prespecified timing boundaries are benchmark rules. Prespecification prevents post-hoc movement of the endpoint but does not itself prove that every action-window boundary is universally correct in all settings.
- The Raw arm received a carefully specified common instruction rather than an unconstrained consumer chat prompt.
- The Guided arm tested a bundled structured-context intervention and does not identify which component produced the observed effect.

- The benchmark did not measure patient comprehension, clinician workflow, adherence, downstream utilization, morbidity, or mortality.
- Secondary structured fields contain 17 semantic consistency queries preserved from the locked scoring record; these do not change the frozen overall categories or primary result and should not be silently recoded after unblinding.

Next validation milestone

The appropriate next step is external replication, not additional statistical manipulation of the same 15 cases. A stronger study should use newly constructed cases, independent reference standards, blinded external clinician scoring, explicit assessment of both under-escalation and over-escalation, and case-clustered inference. A subsequent silent real-world study using de-identified clinical laboratory panels would be needed before making claims about performance in clinical practice.

Conclusions

In a 270-report internal synthetic benchmark, material clinical failure occurred in 18/135 Raw reports and 2/135 Guided reports. The case-balanced difference was -11.85 percentage points (95% paired case-cluster percentile interval, -25.19 to 0.00). The finding is a substantial favorable benchmark signal concentrated in safety-sensitive cases, but it is not proof of real-world effectiveness, patient-safety benefit, or general superiority. Independent external clinical replication is the next required milestone.

Transparency and disclosures

HealthRadar is developing clinical interpretation and evaluation methods and therefore has a potential commercial interest in favorable findings. The authors participated in study design and governance. The clinical scoring process used substantial AI assistance and physician confirmation. AI tools also assisted with drafting, data checking, computational verification, and manuscript preparation. The authors reviewed and take responsibility for the final reported statements and limitations. This study did not evaluate a deployed clinical service and did not measure patient outcomes.

References

1. U.S. Food and Drug Administration. Considerations for the Regulation of Generative AI-Enabled Medical Devices: Discussion Paper and Request for Feedback. 2026.
2. International Medical Device Regulators Forum. Good Machine Learning Practice for Medical Device Development: Guiding Principles. IMDRF/AIML WG/N88 FINAL:2025.
3. Rose EA, Rose SE. HealthRadar Yardstick: Clinical quality and repeatability of large language model interpretations of synthetic laboratory reports. HealthRadar, 2026.
4. HealthRadar. Yardstick 2.0 primary-analysis verification record, HR-YARDSTICK-2-PRIMARY-ANALYSIS-VERIFICATION-001, v1.0.0. 3 October 2026.