CLINICAL AI VENDOR EVALUATION REPORT

Epic Systems Corporation

Epic Sepsis Model (ESM) — retrospective, public record to Aug 2021

Clinical Use Case: Inpatient sepsis prediction / early warning (proprietary penalised logistic regression embedded in the Epic EHR)

Deploying Institution: Retrospective desk evaluation — composite US health-system buyer, vantage point August 2021

Jurisdiction: United States

Evaluation Date: 2026-06-13

Prepared by: Dr Kpakpo Acquaye, MRCEM

EXECUTIVE SUMMARY

Composite Score: 37.0 / 100 - Weak

37.0 / 100

Weak

Recommendation: Do not proceed

One or more domains scored below the safety interlock threshold, so procurement must not proceed until the vendor remediates those deficiencies, regardless of the composite score. Interlock triggered by: Population Validity, Operational Performance. The model-generated narrative for individual domains appears in the domain sections of this report.

SAFETY INTERLOCK TRIGGERED

The following domain(s) scored below 30, triggering a mandatory Do Not Proceed recommendation regardless of composite score:

- Population Validity
- Operational Performance

Supplementary Flags:

- Multi-AI Interaction: Not Addressed
- Vendor Viability: Low Risk

DOMAIN 1: Clinical Evidence Quality - Weight: 25%

Score: 40 / 100 - Weak (D)

The vendor provided retrospective single-site study evidence with no patient-centred outcome data. The independent validation showed significantly lower performance than vendor claims. No prospective validation or peer-reviewed vendor-published external validation was available.

Red Flags:

- No peer-reviewed evidence validating the vendor's own performance claims

Domain score capped at 40

DOMAIN 2: Population Validity - Weight: 15%

Score: 20 / 100 - Insufficient (E)

No demographic composition or subgroup performance analyses were published. The model was deployed at hundreds of hospitals without site-level validation requirements. The single external validation did not address population diversity.

DOMAIN 3: Operational Performance - Weight: 15%

Score: 20 / 100 - Insufficient (E)

The only real-world operational data showed a high alert burden with low positive predictive value. No vendor-published data on missing data handling, latency, fail-safe behaviour, or performance drift monitoring was available.

DOMAIN 4: Workflow Integration - Weight: 15%

Score: 60 / 100 - Limited (C)

The model integrates natively with the Epic EHR, eliminating manual data entry. However, no published data on adoption, override rates, alert response times, or surge behaviour was available.

DOMAIN 5: Regulatory Compliance - Weight: 15%

Score: 40 / 100 - Weak (D)

The model held no FDA clearance or authorisation, and the basis for CDS exemption was unpublished. No post-market surveillance obligations or change-control documentation was publicly available.

Red Flags:

- Tool in active clinical use without formal clearance

Domain score capped at 40

DOMAIN 6: Data Governance & Sovereignty - Weight: 10%

Score: 40 / 100 - Weak (D)

The EHR-native deployment architecture is favourable, but no model-specific data-flow diagrams, hosting terms, or secondary-use terms were publicly disclosed. The proprietary model's computations and data retention were not fully inspectable by customers.

DOMAIN 7: Implementation Maturity - Weight: 5%

Score: 40 / 100 - Weak (D)

The vendor demonstrated substantial implementation reach, but no published requirements for local validation, implementation playbooks, post-go-live audit schedules, or site-level underperformance detection were available.

SUPPLEMENTARY ASSESSMENTS

Multi-AI Interaction Assessment

Status: Not Addressed

No published conflict-resolution protocol, combined alert-burden analysis, or deduplication logic existed.

Vendor Viability Assessment

Status: Low Risk

Epic is a large, long-established EHR vendor, and the ESM was distributed as part of the Epic platform.

Post-Deployment Audit Schedule

Timepoint	Audit Type	Focus
Month 1	Go-live validation	Confirm baseline performance matches vendor claims
Month 3	Early adoption review	Adoption rates, alert volumes, clinician feedback
Month 6	First full audit	Performance drift, nuisance ratio, demographic parity, workflow co
Month 12	Annual governance review	Full re-evaluation against original rubric scores
Ongoing (quarterly)	Monitoring	Automated metrics review with exception reporting

JURISDICTION-SPECIFIC NOTES: United States

Regulatory Compliance

Review FDA 510(k), De Novo, or PMA pathway. Check predicate device validity for AI-specific claims and whether CDS exemption logic is clinician-reviewable.

Data Governance & Sovereignty

Assess HIPAA Business Associate Agreement coverage, state privacy law exposure, data processing locations, and secondary use terms for model retraining.

Implementation Maturity

Confirm post-market monitoring commitments, support coverage across deployment time zones, and escalation paths for clinical safety incidents.

APPENDIX A: SCORING METHODOLOGY

This evaluation uses an 8-domain weighted rubric designed for clinical AI procurement assessment. Each domain is scored 0-100 against explicit tier criteria, then multiplied by its weight to produce a composite score.

Domain Weights (US):

Domain	Weight
Clinical Evidence Quality	25%
Population Validity	15%
Operational Performance	15%
Workflow Integration	15%
Regulatory Compliance	15%
Data Governance & Sovereignty	10%
Implementation Maturity	5%

Score Bands:

Band	Score	Label
A	81-100	Strong
B	61-80	Adequate
C	41-60	Limited
D	21-40	Weak
E	0-20	Insufficient

Composite Score Interpretation:

Composite Score	Interpretation
80-100	Proceed with standard procurement process
60-79	Proceed with conditions - address specific domain weaknesses before deployment
40-59	Significant concerns - vendor must remediate before re-evaluation
20-39	Do not proceed - fundamental deficiencies across multiple domains
0-19	Do not proceed - vendor is not ready for clinical deployment

Safety Overrides:

- Cross-domain safety interlock: any single domain scoring below 30 triggers a mandatory Do Not Proceed recommendation regardless of composite score.

- Automatic red flags: specific findings cap the affected domain at 40.

- Supplementary assessments are reported as qualitative risk indicators alongside the composite score.

This rubric was developed from the perspective of emergency medicine clinical practice, designed to evaluate whether AI tools will survive contact with real-world clinical environments.

APPENDIX B: EVIDENCE SUBMITTED

The following evidence was provided by the evaluator and scored by the model. This is the verbatim input — what you see here is what was scored.

Domain 1: Clinical Evidence Quality

```
Clinical Evidence Quality — public record to August 2021.

The model is a proprietary penalised logistic regression developed by the
vendor, reported as trained on approximately 405,000 patient encounters
across 3 US health systems between 2013 and 2015 [1,4]. Vendor
documentation reported an area under the receiver operating characteristic
curve (AUC) of 0.76-0.83, derived from development/internal data [1,4].

Important scoring guardrail: do NOT trigger a "zero peer-reviewed
publications" red flag for this domain. Wong et al. 2021 is a
peer-reviewed, independently conducted external validation [1], and Habib
et al. is peer-reviewed editorial commentary [3]. The evidence gap is more
specific: as of August 2021, the public record contained no peer-reviewed
prospective validation, no vendor-published peer-reviewed external
validation rebutting Wong et al., and no patient-centred outcome evidence
from the ESM itself (mortality, time-to-antibiotics, ICU transfer).
Scoring guardrail (live-environment flag): do not assert "no evidence of
testing in a live clinical environment" — the model was deployed at scale
and independently evaluated on real hospitalisation data (Wong et al.).
The supportable gap is the absence of published prospective validation and
patient-centred outcome evidence, not an absence of live deployment.

The only independent external validation is Wong et al. 2021 [1]: a
retrospective study at a single large academic medical centre (Michigan
Medicine) covering 27,697 patients and 38,455 hospitalisations. Findings:
AUC 0.63 — substantially below the vendor-reported 0.76-0.83. At the
threshold used operationally, sensitivity was 33% (the model missed 67% of
sepsis cases) with a positive predictive value of 12%. The model identified
only 7% of sepsis cases that clinicians had not already recognised, in part
because some model inputs are entangled with clinician actions (e.g.
antibiotic administration), meaning alerts frequently followed rather than
preceded clinical recognition [1,3].

The vendor publicly disputed the validation, pointing to a smaller Prisma
Health conference abstract, arguing that the Michigan threshold was low,
and stating that the system had helped clinicians save lives [2]. Wong's
lead author responded that the threshold was within Epic's recommended
range and that the Prisma abstract used billing-code-defined sepsis rather
than the clinical criteria used in the independent validation [2]. No
vendor-published peer-reviewed external validation using either definition
was available to rebut the independent findings.

The full model specification (coefficients, feature definitions) was not
publicly available; documentation was provided to customers under
agreement, preventing independent pre-purchase scrutiny [1,3].
```

Domain 2: Population Validity

```
Population Validity — public record to August 2021.

The development cohort was drawn from 3 US health systems (2013-2015) [1,4].
No demographic composition of the training population was publicly
disclosed. No subgroup performance analyses (age, sex, race/ethnicity,
comorbidity strata) had been published by the vendor or independently as of
the vantage date.

The model was deployed at hundreds of US hospitals [1] serving widely
heterogeneous populations, with no published requirement or standard
process for site-level validation prior to go-live, and no published
evidence of performance on populations differing from the development
cohort. The single external validation [1] was itself one academic centre;
generalisability beyond it was unknown in both directions.

The public record is silent on paediatric applicability, performance in
```

immunocompromised populations, and performance across insurance/access strata. Silence is noted as such.

Evaluator assessment note: the public record contains no evidence — as distinct from weak or partial evidence — of performance across population subgroups. No demographic composition, no subgroup analyses by any party, no site-level validation requirement, at a deployment scale of hundreds of heterogeneous hospitals. A complete absence of population-validity evidence for a tool deployed at this scale is an insufficient-evidence posture, not a limited one. The single external validation does not raise this domain: it was one academic centre and reported no subgroup breakdowns.

Domain 3: Operational Performance

Operational Performance — public record to August 2021.

Real-world operational data from the single independent evaluation [1]: at the operational threshold, the model generated alerts on 18% of all hospitalised patients. With a positive predictive value of 12%, approximately 8 patients required clinical evaluation per true positive, creating a documented alert-burden concern raised in the accompanying commentary [3]. Publication wording guardrail: do not describe this as a "93% false positive rate." The directly supported figures are 12% PPV, 8 patients needing evaluation per true positive, alerts on 18% of hospitalisations, and 67% of sepsis cases not identified at the threshold studied [1,3].

No vendor-published data existed on: behaviour with missing or delayed inputs; latency under load; fail-safe behaviour; or performance drift over time. The model in deployment was static, with no published drift monitoring programme or scheduled recalibration as of the vantage date. No published mechanism existed for sites to monitor their own local performance against a reference standard. No demographic performance tracking post-deployment was published.

Uptime and reliability were not separately reported; as a native EHR component, availability tracked the EHR itself (noted as context, not evidence of performance).

Evaluator assessment note: this domain assesses demonstrated operational readiness, which is absent from the public record. The only real-world operational data is independent and adverse — an 18% alert rate at 12% PPV — and the vendor published nothing on drift monitoring, fail-safe behaviour, missing-data handling, or any mechanism for sites to measure their own performance. The existence of independent adverse data does not raise the tier; it documents the operational burden while every operational safeguard remains unevidenced. This is an insufficient posture, not a weak-but-demonstrated one.

Domain 4: Workflow Integration

Workflow Integration — public record to August 2021.

Integration is the model's structural strength. The ESM runs natively inside the Epic EHR: no separate login, portal, or screen; inputs are drawn automatically from the live patient record (vitals, laboratories, medications, demographics); scores recalculate on a scheduled basis during the admission; and alerts surface through the EHR's native alerting (BestPractice Advisory) framework to nursing and clinical staff [1,4]. There is no manual data entry burden.

Limitations in the public record: alerting was not described as role-specific or seniority-adaptive; the operational threshold and alert routing were locally configurable but with no published decision support for choosing them [2]; the 18% alert rate and 12% PPV documented in [1] indicate substantial nuisance-alert exposure for nursing staff in practice; and no surge or mass-casualty mode is described in any public documentation. No published adoption, clinician override-rate, or alert-response-time data existed as of the vantage date. Publication wording guardrail: do not claim "80%+ clinician overrides" for ESM unless a separate source is added.

Evaluator assessment note: the integration architecture is demonstrated and strong — native EHR embedding, no manual data entry — but every human-factors outcome this domain values (adoption, override rates, alert response times, surge behaviour) is unmeasured in the public record, and the documented alert burden falls on nursing staff. Demonstrated architecture with unevidenced use is a limited posture, not an adequate one: strong on plumbing, unproven in human hands.

Domain 5: Regulatory Compliance

Regulatory Compliance — public record to August 2021.

The ESM held no FDA clearance or authorisation; the vendor obtained 510(k) clearances for other products but submitted nothing for sepsis prediction under any pathway [5]. It was distributed as a component of the Epic EHR, implicitly relying on clinical decision support (CDS) treatment outside device regulation. No public documentation articulated the basis for non-device status against the 21st Century Cures Act CDS criteria — in particular the requirement that the clinician be able to independently review the basis of the recommendation, which is in tension with a proprietary model whose coefficients and feature definitions were not publicly disclosed [1,3].

The vendor's intended-use framing (early warning to support clinical judgement) and the operational reality documented in [1] (threshold-driven alerts intended to prompt time-critical sepsis intervention) were not publicly reconciled. Because the tool sat outside the regulated pathway, no post-market surveillance obligations, adverse event reporting requirements, or change-control documentation applied or were published. No quality management system certification specific to the model was publicly cited.

This domain assesses the regulatory posture as a buyer could verify it from the public record; it does not assert that the CDS exemption position was unlawful.

Evaluator assessment note: the vendor held a colorable, industry-standard CDS-exemption position of the kind common across EHR-embedded predictive tools in this period — this is materially different from a vendor with no regulatory theory at all. However, that position was unpublished, unreconciled with the time-critical operational use documented in [1], and unverifiable by a buyer. The posture is weak and undocumented rather than absent or misrepresented.

Scoring guardrail: no regulatory filing of any kind exists for this product, so any red flag premised on a mismatch between marketing claims and a regulatory filing cannot apply. Equally, do not assert as fact that the vendor affirmatively "claimed" a CDS exemption or that the logic was definitively "not clinician-reviewable" — the public record shows the basis for non-device status was unarticulated and unpublished, and the proprietary model was in tension with the Cures Act clinician-review criterion, as described above. The only supportable regulatory red flag is the absence and non-publication of a regulatory position for a tool used to prompt time-critical clinical intervention. Characterise the deficiency as unevidenced and unpublished, never as a proven affirmative misrepresentation.

Domain 6: Data Governance & Sovereignty

Data Governance & Sovereignty — public record to August 2021.

US deployment context: the model was an Epic EHR-native component, so a buyer could reasonably distinguish it from a separate third-party cloud inference product. The public record reviewed here did not provide model-specific data-flow diagrams, hosting terms, telemetry retention terms, or secondary-use terms for the ESM.

However, the public record contains no model-specific disclosures on: whether and how customer data contributed to model development or refreshes; opt-out mechanisms; anonymisation or de-identification audit; or data retention specific to model telemetry. The model's proprietary status meant customers could not fully inspect what the model computed or

retained [1,3]. The public record is largely silent on this domain; the evaluation reflects that silence rather than any documented incident. No data incidents involving the ESM were publicly reported as of the vantage date.

Evaluator assessment note: the EHR-native deployment architecture is structurally favourable compared with a separate external inference vendor, but the public evidence does not prove all data stayed within any one institution-controlled environment, especially for Epic-hosted instances. The deficiencies are therefore unverifiable model-specific specifics, not a documented data incident or documented adverse practice. This is a limited-evidence posture on a comparatively favourable architecture, not a weak or deficient one.

Domain 7: Implementation Maturity

Implementation Maturity — public record to August 2021.

The vendor is the largest EHR supplier to US health systems with a mature, long-established implementation infrastructure, and the ESM had been deployed at hundreds of US hospitals [1] — evidence of substantial implementation reach and repeatable deployment capability.

Material gaps in the public record: there was no published requirement for local validation against the site's own outcome data before go-live; no published implementation playbook for threshold selection, alert routing design, or clinician training specific to the ESM; no published post-go-live audit schedule; and no published process for sites to detect local underperformance. The vendor's public response to the external validation — that hospitals should tune thresholds to local conditions [2] — presumed a local analytic capability that many community deployments lack, and no vendor-provided tooling for that tuning was publicly documented.

Evaluator assessment note: implementation reach is demonstrated at scale, but reach is not maturity. Every maturity mechanism this domain polices — pre-go-live local validation, implementation playbooks, post-go-live audit schedules, site-level underperformance detection — is absent from the public record. Scale without published implementation governance is a weak posture, not a limited one.

Multi-AI Interaction

The public record contains no assessment of the ESM's interaction with other alerting or decision-support systems operating on the same patients. In practice, deploying institutions commonly ran concurrent early-warning scores (e.g. MEWS/NEWS-based tools) and other BPA alerts; no published conflict-resolution protocol, combined alert-burden analysis, or deduplication logic existed as of the vantage date.

Vendor Viability

Vendor viability appears low risk for a US hospital buyer because Epic was already a large, long-established EHR vendor and the ESM was distributed as part of the Epic platform [1]. The stronger claims sometimes made about profitability, private ownership, founding date, or market-share rank are not needed for this retrospective unless separately sourced. Product continuity risk for the ESM is therefore likely lower than for a standalone startup product, although the model's lifecycle (updates, retirement, replacement) remained controlled by the vendor with no published model-specific change-control commitments to customers as of the vantage date.

General Evaluator Notes

None