Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

SepsisShield: Input-Integrity Screening for Trust-Aware Sepsis Early Warning

Israt Jahan Aunika
M.S. Data Science Program, Montclair State University, Montclair, NJ, USA
Corresponding Author: Israt Jahan Aunika, **E-mail**: isratchy433@gmail.com

| ABSTRACT

Early-warning models for sepsis are usually evaluated on clean data, yet bedside data can be wrong in ways a model cannot see, such as a temperature in the wrong units, a monitor artefact or a frozen feed, and the model can remain confident. We present SepsisShield, an end-to-end system that pairs a calibrated sepsis-risk ensemble of five LightGBM models with an independent input-integrity layer that grades the inputs of every patient-hour as HIGH, REDUCED or LOW trust and turns the grade into an action: show, warn or withhold. Only input evidence can withhold a prediction. On 8,068 held-out patients from the PhysioNet/Computing in Cardiology 2019 data, the model reached an AUROC of 0.852 (95% CI 0.840–0.865) and alerted 53.9% of septic patients at least 6 h before onset. Across four simulated accidental fault types, the system flagged or withheld 95.4% (6,481/6,790) of alert decisions that a fault had flipped to a wrong answer (validation replication 94.7%) while withholding 0.46% of clean patient-hours. A component ablation showed that ensembling and calibration alone caught none of these failures and ensemble disagreement caught 4.0%. When only 1% of clean hours can be reviewed, ranking by input trust caught 47.3% of harmful decisions, against 4.4% for model uncertainty; an offline, validation-fitted combination of signals caught 82.2%. Input-integrity evidence thus identified fault-induced unsafe decisions that model-side confidence largely missed, while withholding fewer than 0.5% of clean predictions. Performance remains limited for plausible deliberate edits, frozen feeds and transfer to an unseen hospital.

| KEYWORDS

Clinical decision support; sepsis early warning; data quality; input integrity; trustworthy AI; selective prediction

| ARTICLE INFORMATION

ACCEPTED: 14 October 2026 **PUBLISHED:** 15 November 2026 **DOI:** 10.32996/jcsts.2026.8.16

1. Introduction

1.1 The problem

Sepsis is a dysregulated host response to infection that causes life-threatening organ dysfunction (Singer et al., 2016). A global burden study estimated 48.9 million cases and 11 million sepsis-related deaths in 2017 (Rudd et al., 2020), and delays in effective treatment are associated with lower survival (Kumar et al., 2006). These facts motivate early-warning systems that predict sepsis from routinely collected intensive care unit (ICU) data, and the PhysioNet/Computing in Cardiology Challenge 2019 provided a public benchmark for this task (Reyna et al., 2020).

Accuracy on clean validation data is necessary but not sufficient. Deployed sepsis models have performed worse than expected outside the setting in which they were developed (Wong et al., 2021), and clinical AI is vulnerable to changes in the data it receives (Finlayson et al., 2021). A less discussed failure mode is input corruption at the bedside. Hourly ICU records contain outliers, logically inconsistent blood pressures and omissions (Maslove et al., 2016); unit mix-ups, device artefacts, interface feeds that stop updating and manual edits add to them. A tree ensemble given a body temperature of 98.6 in a Celsius field does not know that the number is impossible: it simply produces a probability, often a confident one. The public data used here already contain such values. Applying the plausibility limits described in Section 6 to all 1,552,210 patient-hours flags 2,548 hours from 1,451 patients with values that cannot be physiologically real, and 1,262 hours with internally inconsistent blood pressure.

SepsisShield therefore asks a second question alongside the usual one. The model asks: *what is this patient's sepsis risk?* An independent integrity layer asks: *can the clinical data behind this prediction be trusted?* The answer to the second question decides

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

what the clinician sees. Put briefly: model confidence asks whether the models agree; input trust asks whether the data deserve to be believed.

1.2 Research question

Can an end-to-end sepsis prediction system detect unreliable clinical inputs and prevent them from silently producing unsafe alert decisions, at an acceptable cost on clean data? We decompose this into four sub-questions, answered in Section 8:

- **RQ1:** Do model-side improvements (ensembling, calibration) or model uncertainty protect against harmful decisions caused by faulty inputs?
- **RQ2:** How many fault-induced harmful decisions does an independent input-integrity layer flag or withhold, and at what cost on clean data?
- **RQ3:** When human review capacity is limited, which signal should decide which predictions are reviewed?
- **RQ4:** Where does the system fail?

2. Literature Review

Sepsis prediction. The 2019 PhysioNet challenge introduced a time-dependent utility score that rewards alerts made between 12 hours before and 3 hours after onset, with the maximum reward 6 hours before onset, and penalises missed, late and false alerts (Reyna et al., 2020). External evaluations of deployed proprietary models have reported substantially lower discrimination than developers reported (Wong et al., 2021), and implementation studies stress workflow integration as much as accuracy (Sendak et al., 2020). An uncertainty-aware model from the challenge estimated the error of a gradient-boosting model and set physiologically implausible values to missing, but did not study corrupted inputs or abstention (Pimentel et al., 2019).

Data quality and validation. Harmonised data-quality frameworks for electronic health records define plausibility and conformance checks for secondary use of clinical data (Kahn et al., 2016), and production machine-learning systems validate incoming data against schemas and expected ranges (Breck et al., 2019). In operational decision-intelligence systems, a pipeline can pass schema, data-quality and drift checks while the meaning of the data has changed, a silent failure that a dedicated multi-layer safety gate can detect (Aunika, 2026). These checks usually run on datasets before training or storage rather than on the inputs to each individual prediction.

Calibration and uncertainty. Isotonic regression is a standard non-parametric calibration method (Zadrozny & Elkan, 2002). Ensembles provide a simple predictive-uncertainty signal through member disagreement (Lakshminarayanan et al., 2017). Uncertainty estimates, however, have been shown to be unreliable for detecting anomalous inputs in medical tabular data (Ulmer et al., 2020), and benchmarks of out-of-distribution detectors on ICU data find that near-distribution anomalies remain difficult (Azizmalayeri et al., 2025).

Selective prediction and deferral. Selective classification lets a model abstain where it is unreliable and studies the trade-off between coverage and risk (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Most such work bases abstention on model uncertainty. Recent clinical triage work defers low-confidence cases to clinicians when deferral lowers expected cost, including under distribution shift (Kwon & Kim, 2026). SepsisShield instead bases abstention on input evidence, and we show empirically why the two are not interchangeable for data faults.

Distribution shift and robustness. Dataset-shift detection has been studied extensively (Rabanser et al., 2019), and clinicians have been urged to treat shift as a patient-safety issue (Finlayson et al., 2021). Robustness benchmarks built from controlled corruptions are established in computer vision (Hendrycks & Dietterich, 2019); our corruption benchmark applies the same idea to ICU time series with clinically motivated fault types.

Explainability. SHAP and its exact tree variant TreeSHAP are widely used to explain tree-ensemble risk scores (Lundberg & Lee, 2017; Lundberg et al., 2020). We use TreeSHAP for risk drivers but also name the integrity check behind every trust decision.

3. Research Gap and Positioning

Prior work offers data-quality checks applied to datasets, uncertainty-based abstention applied to predictions, and corruption benchmarks applied to images. To the best of our knowledge, no prior sepsis early-warning work jointly (a) checks the inputs of every prediction independently of the model, (b) ties the strongest action, withholding, to input evidence rather than model uncertainty, and (c) is evaluated by the number of incorrect alert decisions caused by faults that it intercepts, at a measured cost on clean data. SepsisShield is positioned as an evaluated systems contribution. Its contributions are:

- **C1.** An end-to-end trust-aware pipeline that separates model confidence (agreement between five models) from input trust (evidence from the raw inputs) and maps them to a three-level action, with the rule that only input evidence can withhold a prediction.
- **C2.** A reproducible corruption benchmark on held-out patients with a decision-level safety metric, *dangerous-failure coverage*, which counts only alert decisions that a fault flipped to a wrong answer.
- **C3.** Evidence that model-side improvements and model uncertainty do not protect against faulty inputs, that input integrity does, and that input trust is the better signal for allocating a limited review budget.

- **C4.** An explicit failure analysis, including a pre-specified detector that failed its own success criterion and is reported as a negative result.

Table 1. Positioning against the closest prior work (Y: reported; ~: partial; —: not reported).

Capability	Data-quality frameworks	Uncertainty-based deferral	Corruption benchmarks	Uncertainty-aware sepsis model	SepsisShield
Per-prediction input checks	~	—	—	~	Y
Withholding tied to input evidence	—	—	—	—	Y
Model uncertainty used	—	Y	—	Y	Y (warning only)
Decision-level harm metric	—	—	—	—	Y
Clinical time series	~	~	—	Y	Y
Reported failure modes	—	~	~	—	Y

Sources: data-quality frameworks (Kahn et al., 2016; Breck et al., 2019); uncertainty-based deferral (Geifman & El-Yaniv, 2017; Kwon & Kim, 2026); corruption benchmarks (Hendrycks & Dietterich, 2019); uncertainty-aware sepsis model (Pimentel et al., 2019).

4. System Architecture

SepsisShield has two paths and a decision layer (Figure 1).

- **Prediction path:** raw hourly data → 172 causal features → five LightGBM models → mean probability → isotonic calibration → risk score and alert (threshold 3%) → TreeSHAP risk drivers. Model confidence is the spread of the five models' log-odds.
- **Trust path:** integrity checks run on the raw measurements, before feature engineering, so they do not depend on the model. They grade input trust as HIGH, REDUCED or LOW and record which check fired.
- **Decision layer:** HIGH → show the prediction; REDUCED → show it with a *verify-inputs warning*; LOW → withhold it (the score is kept for audit but not presented as actionable).
- **Advisory context:** a distribution-shift indicator (LOW / MODERATE / HIGH) is displayed beside the decision. It never changes the risk, the alert, the trust level or the decision.

The key design rule is asymmetric. LOW trust, and therefore withholding, is triggered only by evidence from the clinical inputs: physiologically implausible values, impossible blood-pressure combinations, or a coordinated normalising shift across vitals. Ensemble disagreement can lower trust to REDUCED but can never cause a prediction to be withheld. This keeps the strongest action tied to evidence a clinician can check.

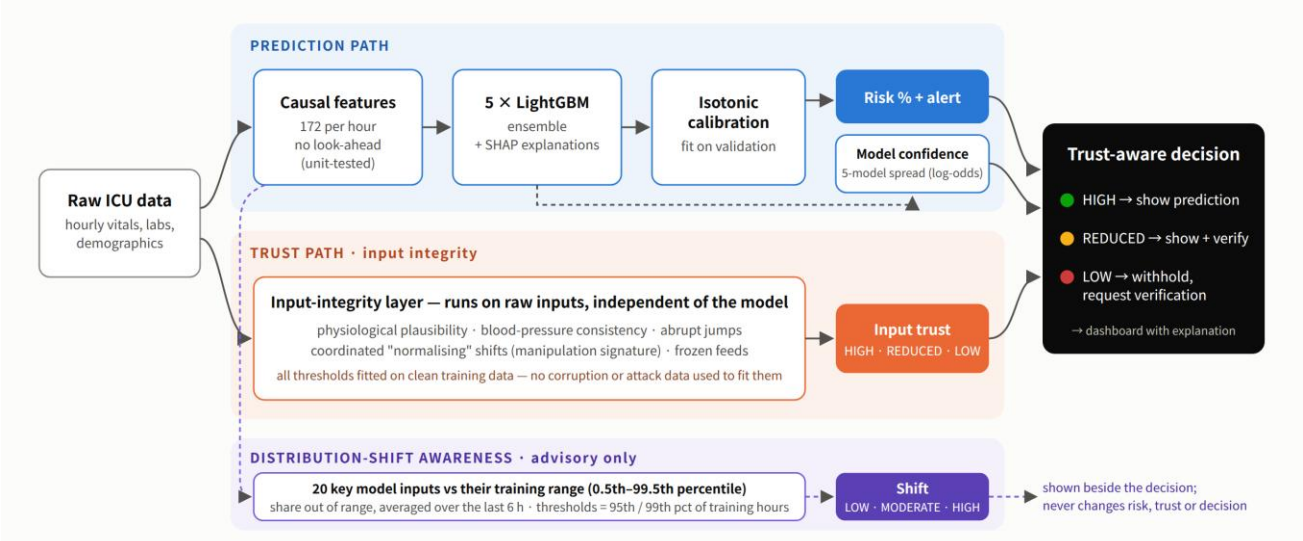


Figure 1. SepsisShield architecture. The prediction path and the trust path run side by side on the same hourly data; a decision layer combines them. The distribution-shift indicator is advisory and never changes the decision.

5. Data and Prediction Model

5.1 Dataset and splits

We use the public PhysioNet/Computing in Cardiology Challenge 2019 training data (Reyna et al., 2020; Goldberger et al., 2000): 40,336 ICU stays from two US hospital systems (20,336 and 20,000 patients), 1,552,210 patient-hours and 2,932 patients who met the challenge's Sepsis-3-based definition. Each hour contains up to 8 vital signs, 26 laboratory values and 6 demographic or administrative variables; most laboratory values are missing during most hours. Patients, never hours, are split 70/10/20 into

training, validation and test sets, stratified by hospital and sepsis status (28,234 / 4,034 / 8,068 patients). The test set has 309,270 patient-hours and 586 septic patients (1.80% of hours positive). The five models were trained on training patients. Isotonic calibration, the alert threshold and the ensemble-disagreement threshold were chosen on validation patients, while the input-integrity rules and the shift reference were derived from training patients only. The held-out test set was not used for any model, threshold or integrity-rule choice.

5.2 Features and model

All 172 features are causal: the value at hour t uses only observations at hours $\leq t$. They comprise last observed values, hours since each variable was last measured, rolling statistics and deltas over 6 and 12 hours, and clinical composites (SIRS score, partial qSOFA, shock index). Missing values are left missing and handled natively by LightGBM (Ke et al., 2017). Five models with different seeds are averaged, and the mean is calibrated with isotonic regression on validation patients. The alert threshold of 3% maximised the challenge utility on validation patients.

To check that the model family is reasonable, we trained random forest, XGBoost (Chen & Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2018) models under the same protocol (Table 2). Gradient-boosted trees dominate; XGBoost lies inside the confidence interval of the SepsisShield ensemble. We kept the LightGBM ensemble because its five members provide the disagreement signal used by the trust layer. The contribution of SepsisShield is not a better risk model; it is what surrounds the risk model.

Table 2. Model families on the held-out test set (8,068 patients, 309,270 hours). Sensitivity: septic patients alerted between 12 h before and 3 h after onset; specificity: non-septic patients never alerted.

Model	AUROC	AUPRC	Utility	Sensitivity	Specificity
SIRS rule (≥2 of 4)	0.635	0.029	0.036	82.1%	26.5%
Logistic regression + isotonic	0.751	0.072	0.284	76.8%	57.4%
Random forest + isotonic	0.839	0.108	0.406	80.2%	68.8%
CatBoost + isotonic	0.839	0.110	0.409	74.7%	77.4%
XGBoost + isotonic	0.853	0.128	0.433	83.4%	68.0%
SepsisShield: 5× LightGBM + isotonic	0.852	0.122	0.425	79.4%	73.8%

6. Input-Integrity Layer

The integrity layer runs five checks on raw hourly measurements. The five input-integrity checks were derived from training patients only:

- **Plausibility:** a value outside physiologically possible limits, deliberately wide ("cannot be real", not "abnormal"; for example, heart rate 20–250 /min, temperature 30–43 °C, systolic pressure 40–280 mmHg).
- **Consistency:** impossible combinations within one blood-pressure reading (diastolic ≥ systolic; mean > systolic + 5 mmHg). Softer discordance that also occurs legitimately is recorded separately.
- **Jump:** an hour-to-hour change larger than the 99.95th percentile seen in clean training data for that vital sign.
- **Coordinated shift:** several vitals moving toward normal together, scored against the patient's own recent baseline; a signature of manual overwriting, thresholded to flag 0.25% of clean training hours.
- **Flatline:** three core vitals identical for at least 8 consecutive hours (a frozen feed or copied-forward values).

Because corrupted values propagate into rolling features, any hard flag keeps the patient at REDUCED trust for 6 hours afterwards. Separately, the ensemble-disagreement warning threshold was selected on validation patients: disagreement above the 99th percentile of validation hours lowers trust to REDUCED. We deliberately report three evidence-based levels and always name the check that fired, rather than a numeric trust score that no outcome data could calibrate. Missing values do not lower trust: most laboratory values are missing during most hours by design, and missingness is itself predictive. Section 8.5 measures what this choice costs.

7. Corruption Benchmark and Evaluation Protocol

Cohort. All corruption experiments use every septic test patient (586) plus 2,500 randomly chosen non-septic test patients (124,773 patient-hours). Faults are injected into the raw data and the full pipeline is re-run: features, models, calibration, integrity checks and shift indicator. No model is retrained. For septic patients, the 10-hour fault window is placed in the clinically decisive period, from 12 hours before to 3 hours after onset.

Faults. Four accidental faults: a thermometer reporting Fahrenheit into a Celsius field; a laboratory unit mix-up between SI and conventional units; a monitor artefact (heart rate 190–240 with systolic pressure 25–45); and a frozen monitor feed. Secondary faults: vitals deliberately overwritten to look normal, the same edit kept physiologically plausible, and ordinary measurement noise as a benign control.

Safety metric. A *dangerous failure* (harmful decision) is a patient-hour inside the fault window in which the fault flipped the alert decision and the new decision is wrong: a timely alert suppressed in a septic patient's useful window, or a new alert outside any

useful window. *Dangerous-failure coverage* is the share of these that are withheld or warned rather than shown silently. The clean-data cost is the share of all 309,270 clean test hours flagged or withheld.

Review budget. For RQ3, each signal ranks all hours, and a fixed share of clean test hours (0.5% to 20%) is sent for review; the outcome is the share of dangerous failures that falls inside the reviewed set. The signals are random review, model uncertainty (ensemble spread), the advisory shift score, input trust, and an offline combined gate: a logistic regression on input trust, ensemble spread and shift score, fitted only on validation patients and evaluated once on the test set. The combined gate is an analysis device for comparing signals, not part of the deployed decision rule or a contribution of the system. Confidence intervals come from 2,000 patient-level bootstrap resamples.

Replication. The integrity checks were derived from training patients; an independent replication used 1,543 validation patients never used to design them.

8. Results

8.1 Clean-data performance

On the held-out test set, the model reached an AUROC of 0.852 (95% CI 0.840–0.865). At the validation-chosen threshold, it alerted 79.4% (75.9–82.5%) of septic patients within the useful window and 53.9% at least 6 hours before onset (median lead time 12 h), and 73.8% of non-septic patients never received an alert. The raw ensemble was already well calibrated (expected calibration error 0.0032); isotonic calibration lowered it to 0.0028 with an unchanged Brier score (0.0166). The false-alert burden is 32.5 new alert episodes per 100 non-septic patient-days at the chosen 3% threshold; a 2% threshold would alert 87.9% of septic patients at 48 episodes, and a 5% threshold 67.2% at 18.

8.2 RQ1: model-side improvements and model uncertainty

Table 3 and Figure 2 add components one at a time. A single model, a five-model ensemble and isotonic calibration each caught 0 of 6,790 dangerous failures: none of them looks at whether the inputs are believable. An ensemble-disagreement warning caught 4.0% (272), because the five models agree with each other on corrupted inputs. Adding the input-integrity checks raised coverage to 95.4% while flagging or withholding 5.1% of clean hours (withholding 0.46%). Also flagging on MODERATE or HIGH shift would have caught 66 more failures (96.4%) but raised clean flags to 9.3%; we judged that trade-off not worth making, so shift remains advisory.

Table 3. Component ablation. AUROC and ECE are on clean held-out data; failures are the 6,790 accidental-fault dangerous failures. The last row is a counterfactual and is not deployed.

Configuration	AUROC	ECE	Dangerous failures caught	Clean hours flagged / withheld
Single LightGBM, raw probability	0.852	0.0031	0 (0.0%)	0.0% / 0.00%
+ five-seed ensemble	0.853	0.0032	0 (0.0%)	0.0% / 0.00%
+ isotonic calibration	0.852	0.0028	0 (0.0%)	0.0% / 0.00%
+ ensemble-disagreement warning only	0.852	0.0028	272 (4.0%)	0.9% / 0.00%
+ input-integrity checks (full system)	0.852	0.0028	6,481 (95.4%)	5.1% / 0.46%
(counterfactual) also flag on shift	0.852	0.0028	6,547 (96.4%)	9.3% / 0.46%

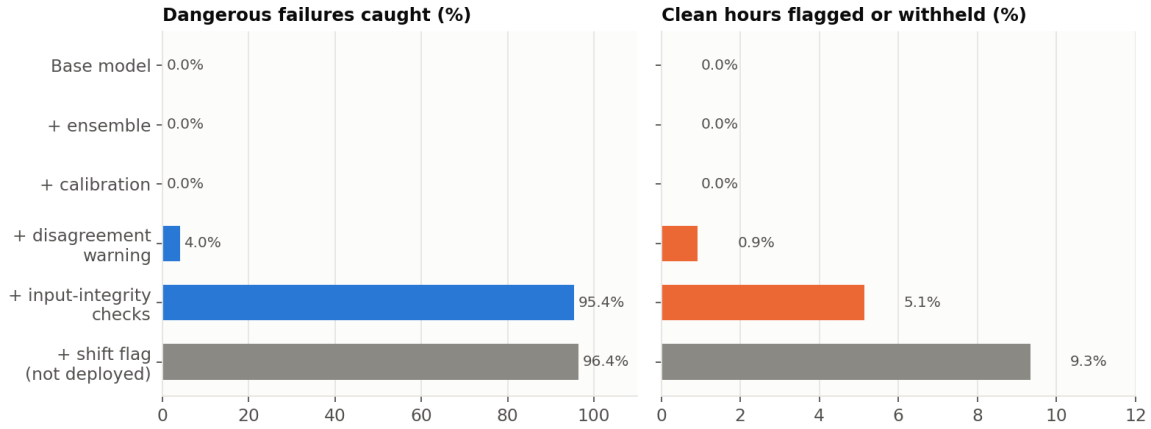


Figure 2. What each component adds. Model-side improvements catch no data-fault failures; the input-integrity checks do almost all of the work. A shift flag would add one percentage point of coverage at the cost of nearly doubling clean flags.

8.3 RQ2: coverage by fault type

Table 4 and Figure 3 break coverage down by fault. Unit errors and monitor artefacts are almost always caught because they produce impossible values. The validation-cohort replication gave 94.7% (3,313 / 3,498) for the accidental faults. Injecting physiologically impossible values (heart rate 900, temperature 70 °C, systolic pressure −20) into 1% of hours produced LOW trust and a withheld prediction at all 1,232 of 1,232 injected hours. On clean data, 12.3% of test patients had at least one withheld hour (median 1 hour).

Table 4. Corruption benchmark on held-out patients. "Warned" = shown with a verify-inputs warning.

Fault (10-hour window)	Dangerous failures	Withheld	Warned	Silent	Coverage
Fahrenheit temperature in Celsius field	4,329	1,609	2,708	12	99.7%
Laboratory unit mix-up	697	200	471	26	96.3%
Monitor artefact	1,255	951	304	0	100.0%
Frozen monitor feed	509	6	232	271	46.8%
All four accidental faults	**6,790**	**2,766**	**3,715**	**309**	**95.4%**
Deliberate edit (vitals made to look normal)	786	66	268	452	42.5%
Deliberate edit kept physiologically plausible	876	33	199	644	26.5%
Benign control: measurement noise	87	2	9	76	12.6%

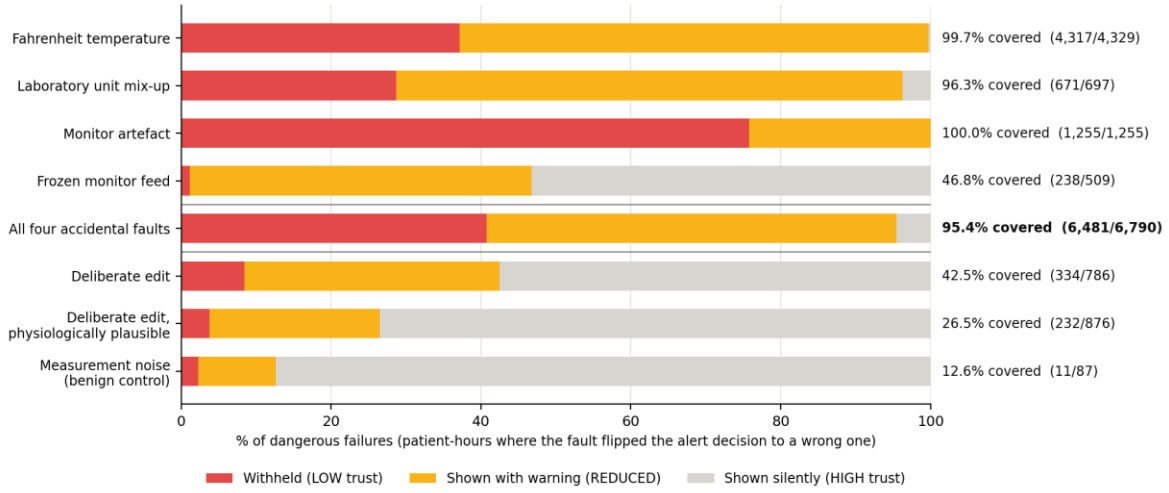


Figure 3. Where fault-induced wrong decisions end up: withheld (LOW trust), shown with a warning (REDUCED) or shown silently (HIGH).

8.4 RQ3: which signal should allocate a limited review budget?

Human reviewers cannot check every prediction. Figure 4 and Table 5 show the share of dangerous failures caught when a fixed share of clean hours can be reviewed. At a 1% budget, model uncertainty caught 4.4% (95% CI 3.5–5.3%), only slightly better than random review (1.1%). Input trust caught 47.3% (45.7–48.8%), 10.9 times as many as model uncertainty (95% CI 8.9–13.4). The validation-fitted combined gate caught 82.2% (79.6–84.7%), 18.9 times as many as model uncertainty (95% CI 15.6–23.3). The gate's advantage over input trust alone holds only at budgets of 2% or less; from 5% upward, the two are equivalent. When the gate was refitted without one fault type, it still outperformed input trust at a 1% budget on the held-out Fahrenheit (94.5% vs 44.3%) and unit-error (48.9% vs 36.4%) faults and matched it on monitor artefacts (80.1% vs 78.5%); neither caught frozen feeds (6.5% vs 6.4%). Because the combined gate uses the advisory shift score, it is reported as an offline analysis, not part of the deployed decision rule.

Table 5. Dangerous failures caught (%) at fixed review budgets (share of clean test hours reviewed), with patient-bootstrap 95% CIs.

Budget	Random	Model uncertainty	Shift score	Input trust	Combined gate
0.5%	0.5	2.1 (1.6–2.8)	30.6 (26.7–34.1)	41.3 (39.7–42.7)	70.4 (68.7–72.2)
1%	1.1	4.4 (3.5–5.3)	41.4 (39.0–43.4)	47.3 (45.7–48.8)	82.2 (79.6–84.7)
2%	2.1	8.5 (7.3–9.7)	51.8 (49.4–55.0)	59.7 (57.9–61.5)	91.0 (89.8–92.3)
5%	5.1	18.5 (16.7–20.2)	63.9 (61.9–65.9)	93.6 (91.0–95.8)	95.4 (94.4–96.2)
10%	10.6	30.9 (28.9–33.2)	91.8 (89.4–93.9)	95.7 (94.8–96.4)	96.4 (95.6–97.1)
15%	15.5	41.1 (38.8–43.4)	93.6 (92.5–94.7)	95.8 (95.1–96.6)	96.8 (96.1–97.5)
20%	20.5	49.3 (47.0–51.6)	94.2 (93.2–95.2)	96.2 (95.4–96.9)	97.1 (96.4–97.7)

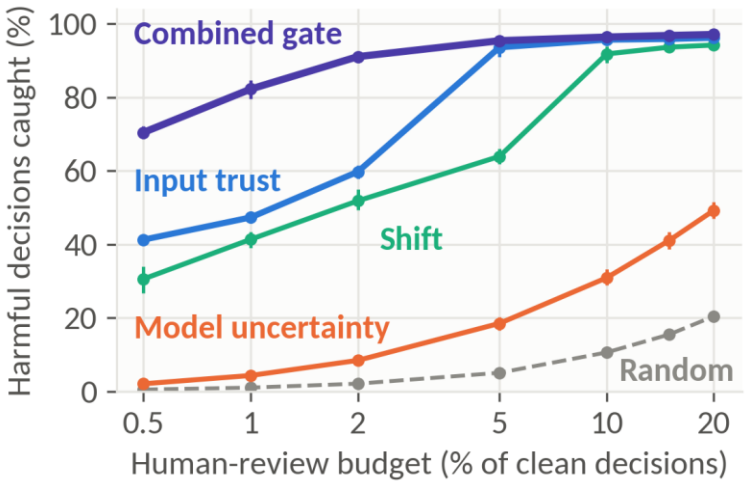


Figure 4. Harmful decisions caught against the human-review budget (log scale), with patient-bootstrap 95% CIs at the seven evaluated budgets.

8.5 RQ4: robustness beyond accidental faults

Missingness and noise. The robustness experiments in this section use the corruption-benchmark cohort of Section 7, which is enriched for septic patients, so its uncorrupted AUROC is 0.811 rather than the 0.852 of the full test set. Discrimination degrades gradually: AUROC fell from 0.811 to 0.791 with 40% of observed values deleted at random, and to 0.802 with noise of half a population standard deviation. Because the trust layer deliberately does not treat missingness as an input fault, the decision changes caused by random missingness are mostly not flagged (6.3% at 40% missingness).

A pre-specified detector that failed. For plausible edits and frozen feeds, we tested a cross-signal consistency detector under a protocol fixed before the test run: regressors predicting each vital from the others, with a pre-specified success criterion of at least 10 percentage points more coverage of plausible edits at no more than 2 points more clean flags. On held-out data, plausible-edit coverage moved from 26.5% to 28.2% while clean flags rose from 5.1% to 5.9%. The detector failed its criterion and was not shipped. A coordinated edit changes several vitals together, so each vital's prediction is built from other, also edited, vitals.

Selective prediction on clean data. Abstaining on the hours with the highest ensemble disagreement lowered the Brier score of the retained hours (0.0166 at full coverage to 0.0121 at 70% coverage, against 0.0167 for random abstention) but lowered AUROC (0.852 to 0.834), one more reason SepsisShield never withholds on disagreement alone.

Distribution shift and transfer. On the test set 3.6% of hours were MODERATE and 1.0% HIGH shift. AUROC on HIGH-shift hours was 0.749 (0.621–0.862) against 0.852 on LOW-shift hours. When the model was trained on data from one hospital and tested on data from the other, AUROC fell to 0.790 (A → B) and 0.775 (B → A), against 0.837 and 0.861 internally; local recalibration improved calibration and specificity but not discrimination.

9. Failure Analysis

- **False alarm from a unit error (caught).** In a held-out patient, a thermometer reporting Fahrenheit for hours 34–41 raised risk at hour 36 from 0.9% to 3.9%, an alert outside the useful window. All five models agreed. The plausibility check fired, trust was LOW and the false alert was withheld at each hour with an impossible reading (hours 35, 36 and 41). In the hours between and after those readings, the corrupted value was carried forward into the features and the risk stayed above the threshold; the 6-hour REDUCED rule showed these false alerts with a verify-inputs warning rather than silently (Figure 5).
- **Hidden alert from an edited chart (caught, but rare).** Vitals overwritten to look normal lowered risk from 3.5% to 2.3%, below the threshold: a confident all-clear. The coordinated-shift detector fired and the prediction was withheld rather than shown as reassuring. Table 4 shows that most edits of this kind are missed.
- **Silent failures.** Of the 6,790 accidental-fault failures, 309 were shown silently; 271 of them came from frozen feeds whose copied-forward values are plausible and change too little for the flatline rule. Random missingness and plausible deliberate edits produce further silent failures.
- **Clean-data false flags.** About half of the clean withholds come from the coordinated-shift detector firing on real patients, likely after rapid improvement following treatment.

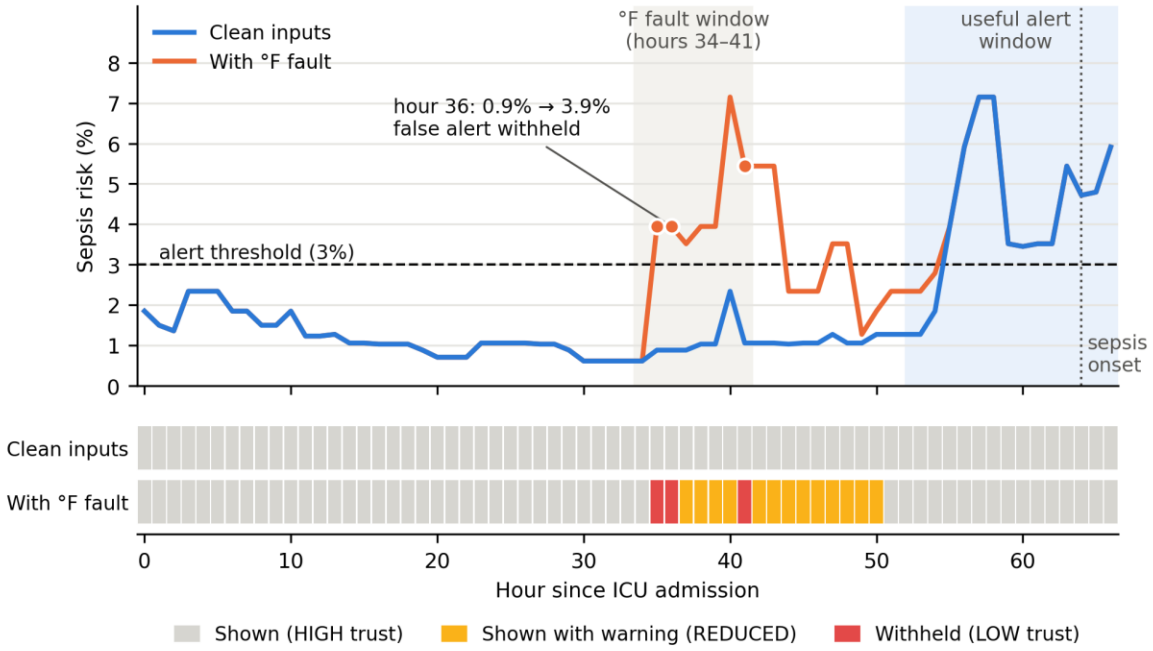


Figure 5. Case study: held-out test patient with a thermometer reporting °F for hours 34–41 (dots: the three °F readings). Top: sepsis risk on clean inputs and with the fault. Bottom: what the clinician sees each hour. The fault creates a false alert 28 h before sepsis onset; it is withheld at the impossible readings and shown with a warning in between, while the clean run shows every hour with HIGH trust.

10. Deployment Considerations

SepsisShield is implemented as an interactive dashboard that loads real held-out patients and shows, for every hour, four cards (sepsis risk, model confidence, input trust and final decision), a clean-versus-corrupted comparison, the TreeSHAP risk drivers, the integrity check that fired, the advisory shift state and the decision rule applied. Users can inject any fault into any demo patient. Every reported number is regenerated from code; an automated test suite pins the demo scenarios to the model's actual outputs. Three operational points follow from the results. First, the alert threshold is a deployment choice; a site would choose its own operating point with local data, and the trust layer works the same at any threshold. Second, the review burden of the trust layer (5.1% of clean hours flagged) is substantial at hourly resolution; Section 8.4 shows that a limited budget is best spent by ranking on input trust rather than model uncertainty. Third, any deployment would need local validation and recalibration, as the cross-hospital results show.

11. Limitations

- **Deliberate manipulation is largely undetected:** 42.5% of alert-changing edits were caught, and 26.5% when edits stayed physiologically plausible. A pre-specified cross-signal detector did not fix this.
- **Missing and noisy data are not flagged;** the decision changes they cause are mostly silent.
- **Generalisation:** AUROC was 0.790 and 0.775 at an unseen hospital, and prevalence or case-mix changes affect discrimination and calibration without being visible to the shift indicator.
- **All faults are simulated.** Their parameters are plausible but not drawn from recorded hospital incidents; real fault rates and mixtures are unknown.
- **Clean-data cost:** 5.1% of clean hours are flagged or withheld and 12.3% of patients have at least one withheld hour.
- **The combined review gate is offline** and depends on the advisory shift score.
- **No prospective or clinician evaluation** has been done. SepsisShield is a research prototype, not a medical device.

12. Conclusion

SepsisShield not only asks what a patient's sepsis risk is; it also asks whether the data behind that prediction can be trusted, and it answers the second question from the clinical inputs, independently of the model. On held-out ICU data, ensembling and calibration alone caught none of the alert decisions that simulated accidental faults turned wrong, and model uncertainty caught 4.0%; the input-integrity layer caught 95.4% while withholding under half a percent of clean predictions. When review capacity is scarce, input trust is a far better guide than model uncertainty. The system has clear limits, namely deliberate edits, frozen feeds, missing data and new hospitals, and those limits are reported alongside the results.

Reproducibility Summary

Table 6. Frozen parameters and verification artefacts.

Item	Value
Patient split (train / validation / test)	28,234 / 4,034 / 8,068 (seeded, patient-level, stratified)
Model	5 LightGBM seeds, 172 causal features, isotonic calibration, threshold 3%
Corruption benchmark (test)	586 septic + 2,500 non-septic patients; 124,773 hours; 6,790 dangerous failures
Validation replication	1,543 patients; 3,498 dangerous failures; coverage 94.7%
Bootstrap	Patient-level; 2,000 resamples (review budget)
Pre-specified negative result	Cross-signal detector: criterion fixed before test run; not met
Automated tests	75 tests passing; demo scenarios pinned to model outputs

Data and Code Availability

The PhysioNet/Computing in Cardiology Challenge 2019 data used in this study are publicly available from PhysioNet. The SepsisShield source code and processed result files are available from the corresponding author upon reasonable request.

Funding: This research received no external funding.
APC Funding: The article processing charge will be paid by the author.
Conflicts of Interest: The author declares no conflict of interest.
Acknowledgments: An earlier version of SepsisShield received the Gold Award at the Global Innovation Build Challenge V2 (2026), Track 02: Medical Technology & Finance.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

Azizmalayeri, M., Abu-Hanna, A., & Cinà, G. (2025). Unmasking the chameleons: A benchmark for out-of-distribution detection in medical tabular data. International Journal of Medical Informatics, 195, Article 105762. https://doi.org/10.1016/j.ijmedinf.2024.105762

Aunika, I. J. (2026). MIPODIS: A migration-ready operational decision-intelligence framework for safe data-source transitions. Journal of Computer Science and Technology Studies, 8(9), 80–93. https://doi.org/10.32996/jcsts.2026.8.9.9

Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data validation for machine learning. In Proceedings of Machine Learning and Systems (MLSys 2019). https://mlsys.org/Conferences/2019/doc/2019/167.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283–286. https://doi.org/10.1056/NEJMc2104626

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems 30 (pp. 4878–4887). https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=HJz6tiCqYm

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., Liaw, S.-T., Hamilton-Lopez, M., Meeker, D., Ong, T. C., Ryan, P., Shang, N., Weiskopf, N. G., Weng, C., Zozus, M. N., & Schilling, L. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. eGEMs, 4(1), Article 1244. https://doi.org/10.13063/2327-9214.1244

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30 (pp. 3146–3154). https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine, 34(6), 1589–1596. https://doi.org/10.1097/01.CCM.0000217961.75225.E9

Kwon, H., & Kim, D.-J. (2026). Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift. Scientific Reports, 16, Article 10016. https://doi.org/10.1038/s41598-026-40637-w

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems 30 (pp. 6402–6413). https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Maslove, D. M., Dubin, J. A., Shrivats, A., & Lee, J. (2016). Errors, omissions, and outliers in hourly vital signs measurements in intensive care. Critical Care Medicine, 44(11), e1021–e1030. https://doi.org/10.1097/CCM.0000000000001862

Pimentel, M. A. F., Mahdi, A., Redfern, O., Santos, M. D., & Tarassenko, L. (2019). Uncertainty-aware model for reliable prediction of sepsis in the ICU. In 2019 Computing in Cardiology (CinC) (pp. 1–4). https://doi.org/10.23919/CinC49843.2019.9005942

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems 31 (pp. 6638–6648).
https://proceedings.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In Advances in Neural Information Processing Systems 32 (pp. 1396–1408).
https://proceedings.neurips.cc/paper_files/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html

Reyna, M. A., Josef, C. S., Jeter, R., Shashikumar, S. P., Westover, M. B., Nemati, S., Clifford, G. D., & Sharma, A. (2020). Early prediction of sepsis from clinical data: The PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine, 48(2), 210–217.
https://doi.org/10.1097/CCM.0000000000004145

Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., Colombara, D. V., Ikuta, K. S., Kissoon, N., Finfer, S., Fleischmann-Struzek, C., Machado, F. R., Reinhart, K. K., Rowan, K., Seymour, C. W., Watson, R. S., West, T. E., Marinho, F., Hay, S. I., . . . Naghavi, M. (2020). Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the Global Burden of Disease Study. The Lancet, 395(10219), 200–211. https://doi.org/10.1016/S0140-6736(19)32989-7

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., Kester, K., Sandhu, S., Corey, K., Brajer, N., Tan, C., Lin, A., Brown, T., Engelbosch, S., Anstrom, K., . . . O'Brien, C. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. JMIR Medical Informatics, 8(7), e15182. https://doi.org/10.2196/15182

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8), 801–810. https://doi.org/10.1001/jama.2016.0287

Ulmer, D., Meijerink, L., & Cinà, G. (2020). Trust issues: Uncertainty estimation does not enable reliable OOD detection on medical tabular data. In Proceedings of the Machine Learning for Health NeurIPS Workshop (Proceedings of Machine Learning Research, Vol. 136, pp. 341–354).
https://proceedings.mlr.press/v136/ulmer20a.html

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 694–699). https://doi.org/10.1145/775047.775151