



Vega Health Benchmarking Insights

Retrospective Evaluation of Adult Inpatient Deterioration Models

For health system leaders struggling with high rates of inpatient mortality and ICU transfers, evaluating AI/ML models that predict deterioration on local patient data can build clinician confidence and reveal operational tradeoffs for clinical workflows.

September 30, 2026 | Insights Lab

By Chris Provan, Joanne Kim, Aden Klein, Merritt Cahoon, and Mark Sendak



Sepsis and acute respiratory failure are among the most common drivers of acute inpatient deterioration, carrying an in-hospital mortality of 20-30% and contributing to 34-52% of all in-hospital deaths.¹

Chart-based scoring systems like the National Early Warning Score (NEWS) 2 can give clinicians a way to evaluate patient stability and identify at-risk patients. Health systems, such as Duke Health,² Kaiser Permanente,³ and the University of Michigan,⁴ have also developed and deployed sophisticated AI/ML algorithms to predict deterioration risk and support clinician decision-making. EHR vendors also make predictive deterioration models widely available⁵ to clinical leaders.

With so many options available, clinical leaders need a way to evaluate the optimal AI/ML model and deployment approach for their care teams and patient populations. Retrospective model evaluations provide a powerful tool to establish local performance before any model is deployed in a live clinical workflow. Local retrospective evaluations generate the evidence that health system leaders need to select and configure AI for clinical use.

Evaluation Design

A multistate health system compared the performance of four different models across 118,063 historical patient encounters leveraging the Vega Health Platform. Applying a consistent methodology allowed leaders to compare each model's positive predictive value⁶ and sensitivity, including model performance across patient subgroups.⁷ The evaluation first compared each model's statistical performance, including the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), sensitivity, and positive predictive value. These results were then interpreted as clinical and operational measures, such as event detection rate and alert burden, to reflect how the model is likely to impact workflows and responsiveness when integrated into the clinical environment.

Exhibit 1: Models included in the evaluation

Model	Type	Origin	Input Data Elements	Output
Vendor Deterioration Index	Predictive: Ordinal Logistic Regression Model	Developed by health IT vendor	Uses 17 features including demographic information, lab results, vital signs, and nursing assessments	0-100 relative risk score
National Early Warning Score (NEWS) 2 Scoring System	Rule-based: Aggregate Scoring System	UK Royal College of Physicians	6 physiological parameters and presence of supplemental oxygen	0-20 aggregate score
Duke Health Deterioration Index	Predictive: Light Gradient-Boosting Machine Model	Developed by Duke Health	Uses demographics, comorbidities, labs, vitals, medication administrations, and order	0-100% relative risk score
Duke Deterioration Phenotypes	Rule-based: Flags (yes/no) based on threshold of criteria	Developed by Duke health	Labs, vitals, medication administrations, orders, supplemental oxygen	Yes/No determination for one or more of the six phenotypes

Results from the retrospective cohort indicated that the Duke Deterioration Index model outperformed both a widely used vendor deterioration index model and the NEWS 2 Scoring System as a predictive instrument. Importantly, both accuracy in deterioration predictions and the number of deterioration events captured could be improved by combining a predictive model (e.g., Duke Deterioration Index) with a rules-based score (e.g., NEWS 2).

Research Methodology

This model evaluation included baseline and comparative performance data for a widely used deterioration index, NEWS 2 Scoring System, and the Duke Health Deterioration Index Model (independently and in conjunction with the Duke Health Deterioration Phenotypes or NEWS 2). Notably, this is the first evaluation of the Duke Deterioration Index model on retrospective patient encounters outside of the Duke Health System.

The historical patient encounter cohort for this evaluation included 118,063 individual encounters occurring between 05/01/2024 and 04/30/2026 across five hospitals. Vega Health generated hour-by-hour risk scores for each individual patient across encounters for the Duke Health Deterioration Index Model and NEWS 2 Scoring System, then compared corresponding alerts from the existing deterioration index as recorded contemporaneously in the EHR.

Deterioration events were defined as either unplanned escalation to an Intensive Care level of care within 12 hours or death within 24 hours following the deterioration alert or prediction. Encounters were classified as either deterioration events or non-events.

This retrospective evaluation first compared model performance across key statistical measures, then analyzed the accuracy, actionability, comprehensiveness, and end-user burden associated with the use of each model in clinical workflows.

Statistical Measures of Model Performance: Models and scores were evaluated across threshold-independent metrics, including Area Under the ROC (AUROC) and area under the precision-recall curve (AUPRC). The evaluation also included each model's sensitivity, positive predictive value, and the number needed to evaluate at equivalent alert rates and sensitivities. Finally, the evaluation compared subgroup-specific performance, identifying potential bias or performance limitations.

Workflow Implications: This analysis also includes insights on how model statistical performance may affect implementation when deployed into clinical workflows, including alert burden, false positive rates, and configurations that could optimize model performance above the baseline.

Statistical Model Performance Comparisons

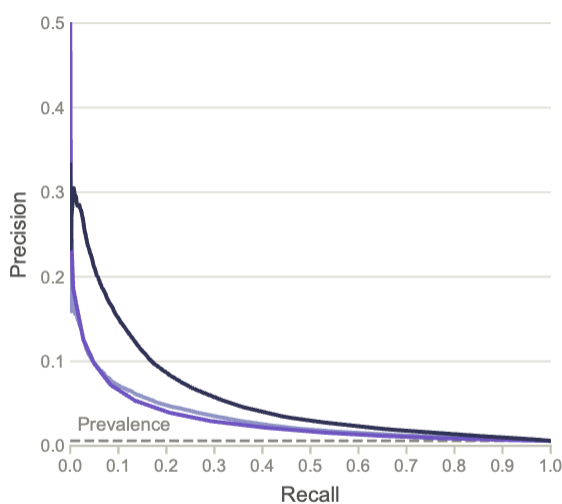
This evaluation first compared each model's ability to distinguish between patients who may or may not experience deterioration across certain thresholds. The Duke Deterioration Index model outperformed both the widely-used vendor deterioration index model and the NEWS 2 Scoring System as the best overall discriminator.

The Receiver Operating Characteristic (ROC)⁸ and Precision-Recall (P-R)⁹ curves indicated that the Duke Deterioration Index model correctly identified a higher share of deterioration events with a higher share of accurate alerts than the other two predictors across all possible thresholds.

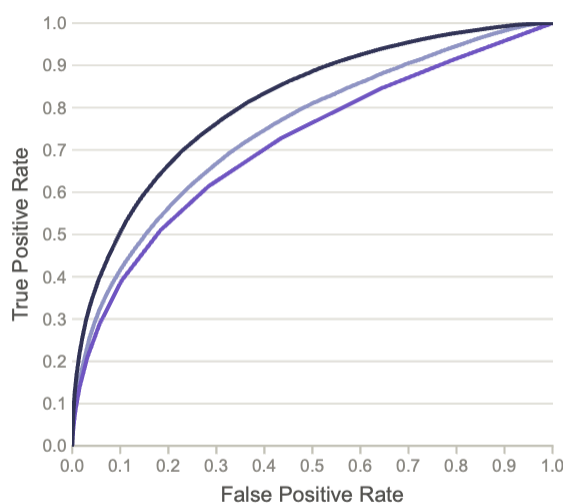
The greater predictive performance of the Duke Deterioration Index model on this cohort was highly statistically significant when compared to NEWS 2 and the vendor deterioration index model, with p-values under 0.001 when comparing the Area Under the Curve (AUC).

Exhibits 2 & 3: Threshold-Independent Model Performance

Precision-Recall Curves



Receiver Operating Characteristic Curves



Duke Deterioration Index
AUPRC = 0.057 | AUROC = 0.812

Widely Used Vendor Deterioration Index
AUPRC = 0.032 | AUROC = 0.748

NEWS 2
AUPRC = 0.025 | AUROC = 0.714

The Operational Reality – Event Detection Rate & Alert Burden

The predictive ability of an AI model is only part of the picture. When health systems deploy a model, they must also understand how it will affect clinical workflows. The retrospective evaluation also compared the number of alerts each model would have generated and the percentage of total deterioration events that the model would have predicted.

The widely used vendor deterioration index model and NEWS 2 Scoring System were evaluated at their default alerting thresholds, estimating alert frequencies and the deterioration frequency corresponding to medium and high alert levels for each tool. The Duke Deterioration Index model does not specify default risk score thresholds to trigger an alert, so the model was evaluated at alert thresholds corresponding to a comparable 50% (medium) and 30% (high) sensitivity level.

For the widely used vendor deterioration index model, default risk score thresholds to trigger an alert captured over 60% of total deterioration events; however, the volume of alerts was high, and only a small percentage of alerts accurately predicted a true deterioration event.

Exhibit 4: Widely Used Vendor Deterioration Index Threshold-Level Performance

Threshold Category	Range	Percent of Scores	Percent Captured Events	Alerts per Patient-Day	Deterioration Prevalence
Low (Green)	<0.45	76.2%	39.6%	-	0.3%
Medium (Yellow)	0.45-0.59	18.5%	29.8%	4.4	1.0%
High (Red)	0.6+	5.3%	30.6%	1.3	3.4%

Similarly, the NEWS 2 Medium and High thresholds captured over 50% of deterioration events but significantly over-alerted, resulting in a low percentage of positive alerts out of a high overall alert volume.

Exhibit 5: NEWS 2 Score Threshold-Level Performance

Threshold Category	Range	Percent of Scores	Percent Captured Events	Alerts per Patient-Day	Deterioration Prevalence
Low	0-4	77.2%	45.5%	-	0.4%
Low-Medium	3 in single dimension	5.9%	5.0%	1.2	0.6%
Medium	5-6	12.4%	21.9%	3.0	1.1%
High	7+	5.5%	27.7%	1.3	3.1%

At thresholds configured to capture a similar share of deterioration events (i.e., fixed sensitivity), the Duke Deterioration Index model alerted at a lower rate than the widely used vendor deterioration index model and NEWS 2. Additionally, a higher percent of alerts triggered by the Duke Deterioration Index model accurately predicted a true deterioration event (i.e., improved positive predictive value). Enabling a 4-hour alarm “snooze” further reduced alert burden.

Exhibit 6: Duke Deterioration Index Threshold-Level Performance

Threshold Category	Range	Percent of Scores	Percent Captured Events	Alerts per Patient-Day	Deterioration Prevalence
Low	<0.0065	89.5%	50.0%	-	0.3%
Medium	0.0065-0.014	7.2%	20.0%	1.7 (0.6 w/ 4-hour alarm snooze)	1.7%
High	0.014+	3.3%	30.0%	0.8 (0.3 w/ 4-hour alarm snooze)	5.5%

Head-to-Head Performance

To capture how the models performed head-to-head, the Duke Deterioration Index model's thresholds were adjusted to match the widely used vendor model and NEWS 2's alert rates¹⁰ and sensitivity levels directly, isolating differences in model accuracy from differences in configuration.

Exhibit 7: Matched Head-to-Head Comparison

	Matched Alert Rate (High alerts)	Matched Sensitivity (High alerts)
Widely Used Vendor Deterioration Index Model vs Duke	Duke Deterioration Index captures 55% more deterioration events	Duke needs 39% fewer alerts; 76% fewer with 4-hour snooze
NEWS 2 vs Duke	Duke Deterioration Index model captures 39% more deterioration events	Duke needs 50% fewer alerts; 80% fewer with 4-hour snooze

Optimizing Outcomes with Combined Algorithm Deployment

At Duke Health, the predictive Duke Deterioration Index model was deployed with the rules-based Duke Deterioration Phenotypes to support clinician workflows (see Duke Health Case Study for additional insights)¹¹. For systems that want to prioritize explainability for bedside nurses, this approach also has the ability to improve event detection rate and increase clinician confidence.

Duke Health Case Study

At Duke Health, the Deterioration Index model alerts are paired with notifications about whether a patient meets the criteria for one of six physician-developed deterioration phenotype flags. The predictive deterioration index model identifies early deterioration risk before visible clinical signs appear, while the phenotype flags surface specific, observable indicators tied to known deterioration pathways and complications. The combination provides clinicians with both an early warning and a concrete clinical reason behind it.

Exhibit 8: Duke Deterioration Phenotypes

Phenotype	Definition	Phenotype	Definition
Hypotension	Patient has either: 1) Two SBP < 90 mmHg within 6 hours 2) Two MAP < 65 mmHg within 6 hours	Vasoactive Medications	New administration or increased dose of at least one vasopressor or inotrope
End Organ Dysfunction	Patient has any of the following criteria: 1) Acute Renal Insufficiency (↑ Creatinine ≥ 0.3mg/dl or Creatinine ≥ 1.5x baseline) – exclusions: ESRD or active HD 2) Lactate > 2.0 mmol 3) AST or ALT > 5x ULN 4) Total Bilirubin > 2.0 mg/dL	Respiratory decline	Patient experiences any indicator of respiratory decline: 1) Increase in O2 (> 2L) within 6 hours 2) O2 saturation below 91% 3) PaO2 < 60mmHg 4) PaO2 decrease by 10mmHg
Hypoperfusion	Patient meets both: 1. Hypotension criteria AND 2. Develops End Organ Dysfunction criteria within 24 hours	Respiratory intervention	Patient has a new oxygen requirement (>6L via NC) or escalating level of respiratory support

This evaluation recreated Duke Health’s combined deployment approach to validate whether layering predictive model alerts with rule-based scores could improve the overall predictive performance. When multiple phenotype flags were present, the overall prevalence of deterioration events for each Duke Deterioration Index alert threshold increased significantly.

Exhibit 9: Deterioration Prevalence by Combined Alert Level & Phenotypes

	0 Phenotypes	1 Phenotype	2 Phenotypes	3+ Phenotypes
Duke: Low	0.3%	0.5% +20 bps vs. 0 Phenotypes	1.0% +70 bps vs. 0 Phenotypes	3.0% +270 bps vs. 0 Phenotypes
Duke: Medium	1.3%	2.1% +80 bps vs. 0 Phenotypes	3.0% +170 bps vs. 0 Phenotypes	6.9% +560 bps vs. 0 Phenotypes
Duke: High	3.0%	6.3% +330 bps vs. 0 Phenotypes	8.8% +580 bps vs. 0 Phenotypes	17.2% +1,420 bps vs. 0 Phenotypes

The retrospective evaluation also compared whether layering predictive model alerts on top of the NEWS 2 Scoring System could improve predictive performance over the baseline.

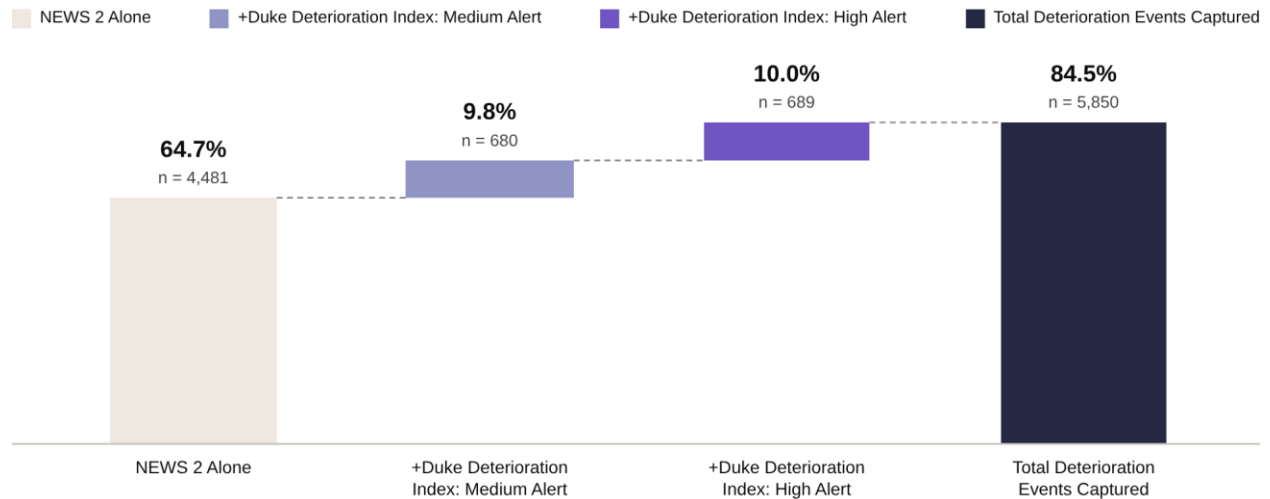
When both the NEWS 2 Scores and the Duke Deterioration Index model generated an alert, the positive predictive value was much higher (8.7%) than for either tool alone (5.5% for Duke Deterioration Index model alone, 3.1% for NEWS 2 alone).

Exhibit 10: Combining NEWS 2 with Duke Deterioration Index Alerts Increases Predictive Value

	Overall Prevalence	Duke: Low	Duke: Medium	Duke: High
NEWS 2: Low	0.4%	0.3%	1.3% +90 bps vs. NEWS 2 alone	3.0% +260 bps vs. NEWS 2 alone
NEWS 2: Medium	1.1%	0.6%	1.9% +80 bps vs. NEWS 2 alone	5.1% +400 bps vs. NEWS 2 alone
NEWS 2: High	3.1%	1.1%	2.8%	8.7% +560 bps vs. NEWS 2 alone

Furthermore, combining both NEWS 2 and Duke Deterioration Index model captured more deterioration events (i.e., increases sensitivity). Over the two-year period, 35% of deterioration events never triggered a medium or high alert in NEWS 2, but the Duke Deterioration Index model alerts may have captured an additional 19.8% of deterioration events (56% of missed events).

Exhibit 11: Combining NEWS 2 with Duke Deterioration Index Alerts Captures More Deterioration Events

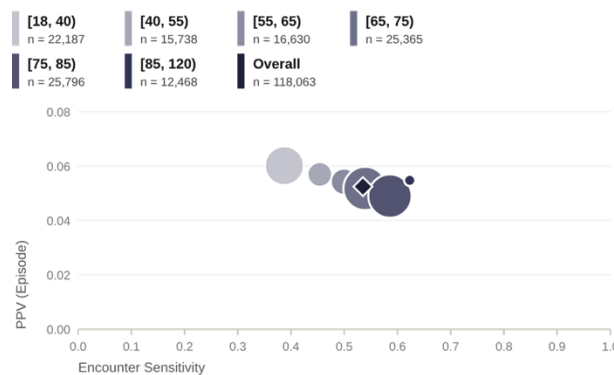


Identifying Risk & Bias

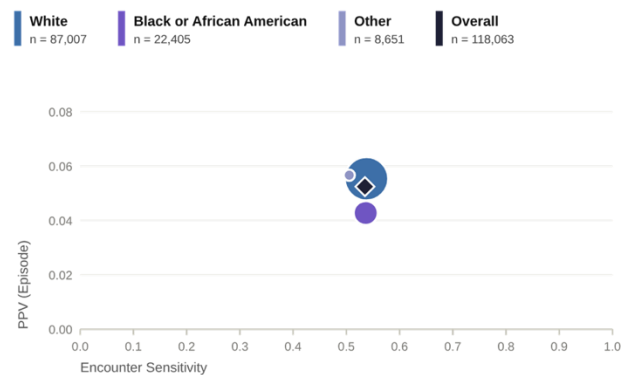
Retrospective evaluations also provide health system leaders with insights into where a model should not be deployed. When models were evaluated for each population subgroup – age band, biological sex, preferred language, and race – performance was typically within expected bounds.

Exhibit 12: Duke Deterioration Index Subgroup Performance

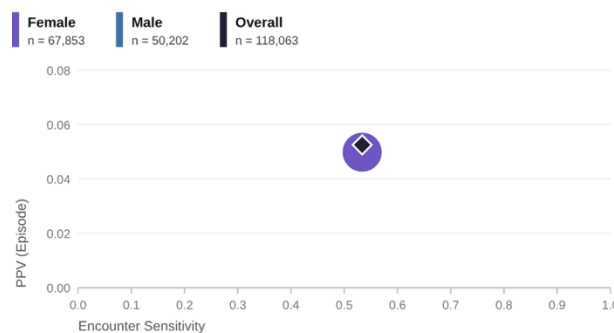
Model Performance by Age



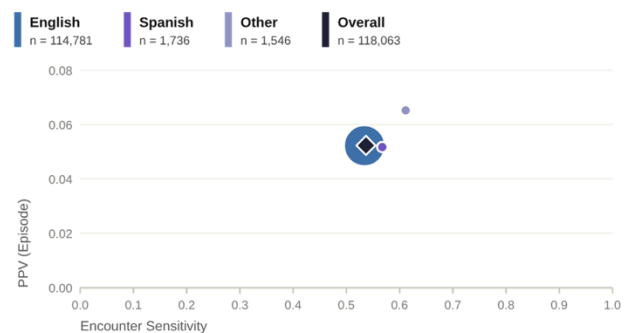
Model Performance by Race



Model Performance by Sex

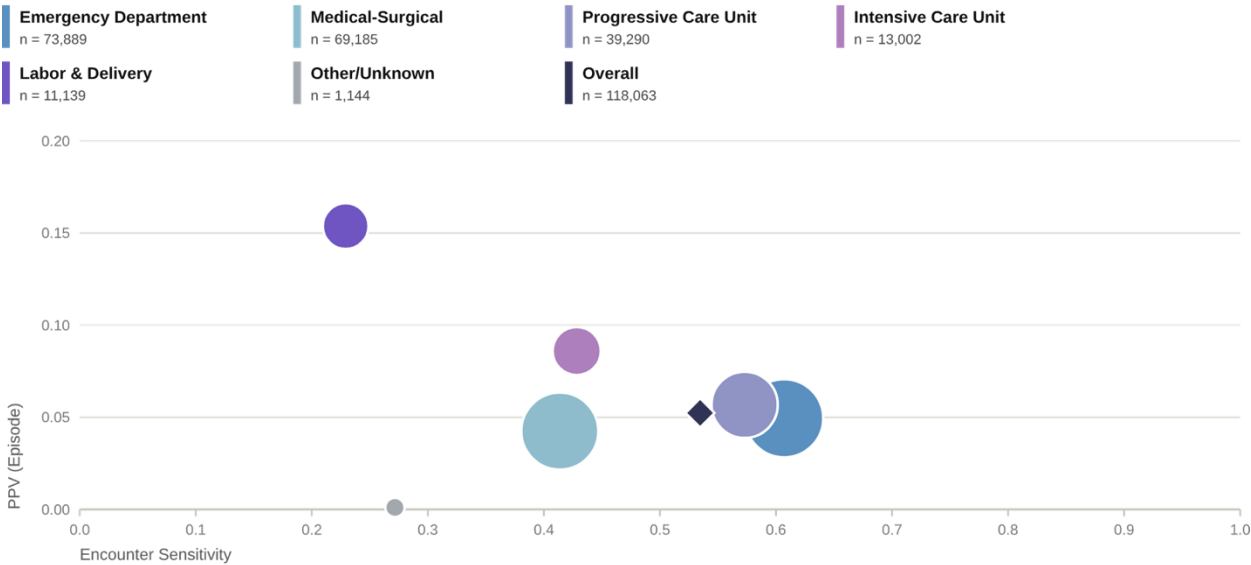


Model Performance by Language



However, model performance was below expected bounds in one specific hospital context. In Labor & Delivery, the evaluation identified major performance discrepancies. This provided clinical leaders with the information necessary to determine which departments should be in scope for model implementation.

Exhibit 13: Duke Deterioration Index Performance by Department



Make AI Deployment Decisions with Confidence

The Vega Health Platform provided a multistate health system with the capabilities necessary to quickly conduct a head-to-head model comparison and identify the optimal implementation approach (e.g., combined deployment) on a local patient cohort. Insights like those in this report de-risk AI model selection, deployment, and change management, making it easier to pick the best-performing model and anticipate how it will operate within clinical workflows.

Note to the Reader: The Duke Deterioration Index model and phenotypes were included in the head-to-head comparison through a distribution partnership between Duke Health and Vega Health. The widely used vendor deterioration index model and NEWS 2 Scoring System were included to help inform a health system partner’s AI deployment. For other health systems that may be interested in a similar model evaluation, there will likely be variability in the results based on the system’s patient population and workflows, which is why rigorous, site-specific benchmarking matters.

About the Authors

Chris Provan leads AI Evaluation & Monitoring for Vega Health. **Joanne Kim**, Director of Product & Services, helps Vega Health partners leverage AI investments to drive clinical and operational value. **Mark Sendak** is the Co-Founder and CEO of Vega Health.

The authors would like to thank Aden Klein and Merritt Cahoon for contributions to this report.

¹ Romero-Brufau, S., Gaines, K., Nicolas, C. T., Johnson, M. G., Hickman, J., & Huddleston, J. M. (2019). The fifth vital sign? Nurse worry predicts inpatient deterioration within 24 hours. *JAMIA open*, 2(4), 465–470. <https://doi.org/10.1093/jamiaopen/ooz033>

² Duke Institute for Health Innovation. 2026. DIHI Impact Report, Volume 26. Duke Institute for Health Innovation.

³ Martinez VA, Betts RK, Scruth EA, Buckley JD, Cadiz VR, Bertrand LD, et al. The kaiser permanente northern california advance alert monitor program: an automated early warning system for adults at risk for in-hospital clinical deterioration. *Jt Comm J Qual Patient Saf*. 2022;48(8):370–5. <https://doi.org/10.1016/j.jcjq.2022.05.005>

⁴ Jacokes P. MCURES [Internet]. AI & Digital Health Innovation. 2023 [cited 2026 Sept 15]. Available from:

<https://aidhi.umich.edu/emerging-products-blog/predicting-c-difficile-infection-cdi-j6nt9>

⁵ Singh, K., Valley, T. S., Tang, S., Li, B. Y., Kamran, F., Sjoding, M. W., Wiens, J., Ottes, E., Donnelly, J. P., Wei, M. Y., McBride, J. P., Cao, J., Penzoza, C., Ayanian, J. Z., & Nallamothu, B. K. (2021). Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Annals of the American Thoracic Society*, 18(7), 1129–1137. <https://doi.org/10.1513/AnnalsATS.202006-698OC>

⁶ True Positives as a percent of positive *predictions* on an hourly prediction basis (i.e., What % of hourly alerts accurately predict an upcoming deterioration event?)

⁷ True positives as a percent of positive *outcomes* (deteriorations) on an hourly prediction basis (What % of hours that *should* predict a deterioration *do* predict a deterioration?)

⁸ The Receiver Operating Characteristic curve compares the share of alerts that accurately capture deterioration events to the share of alerts that result in false positives, independent of the frequency of deterioration events.

⁹ The Receiver Operating Characteristic curve compares the share of alerts that accurately capture deterioration events to the share of alerts that result in false positives, independent of the frequency of deterioration events.

¹⁰ Positive predictions as a percent of total predictions (What % of hourly predictions result in an alert?)

¹¹ Duke Institute for Health Innovation. 2026. *DIHI Impact Report, Volume 26*. Duke Institute for Health Innovation.