Standard EWS evaluations pool observations across time, yielding misleading estimates of net benefit and utility. Our calendar-day framework provides clinically meaningful estimates.

A Clinically Intuitive Approach to Evaluate Performance of Early Warning Scores

Peng Wu, MS[1], Kyle Carey, MPH[2], Sierra Strutz, MS[1], Majid Afshar, MD, MS[1], Dana Edelson, MD, MS[2], Matthew Churpek, MD, MPH, PhD[1], William F. Parker, MD, PhD[2], Anoop Mayampurath, PhD[1]

[1]University of Wisconsin-Madison, [2]University of Chicago

Introduction

Early Warning Scores (EWS): widely used to detect deterioration in hospitalized patients.[1-4]

Evaluation Metrics: Discrimination (e.g., AUC) do not reflect clinical trade-offs of benefit vs. harm; Net benefit and utility better capture clinical impact.[5]

Problem: Standard evaluations pool observations across years, assuming patients who are NOT hospitalized together share the same care resources.

→ misleading estimates of net benefit and utility.

Objective: Propose a new calendar-day framework that evaluates patients hospitalized concurrently, yielding realistic and clinically meaningful estimates of net benefit and utility.

Methods: Data & Framework

Data: Retrospective cohort of adult hospital admissions, University of Chicago (2009–2020); split into derivation (2009–2014) and internal validation (2015–2020) cohorts.

EWS Models: MEWS, NEWS, Logistic Regression (LogReg), Recurrent Neural Network (RNN), XGBoost (XGB).

Outcome: Clinical deterioration (unplanned ICU transfer or death within 24h).

Evaluation Framework

Standard pooling: All observations combined to get net benefit and utility.

Calendar-day evaluation:

- Patients present at 8 AM each calendar day (concurrent daily cohorts). See Figure 1.
- Outcomes are determined within the following 24h.
- Net benefit and utility are represented as median (IQR) across daily estimates.

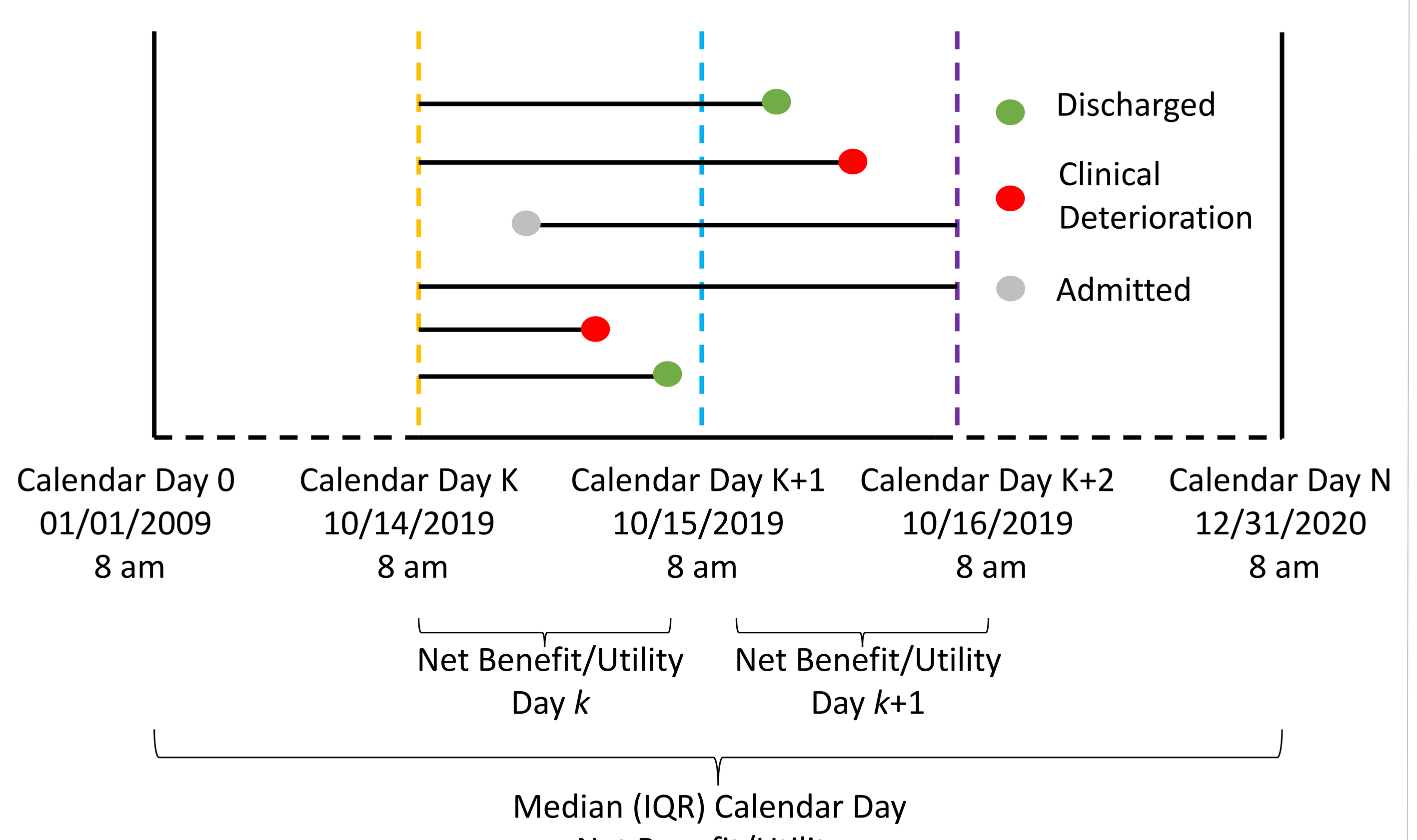


Figure 1. Illustration of calendar-day net benefit/utility at 8 AM.

Methods: Metrics

Metrics:

- Net Benefit: decision-analytic benefit vs. harm.
- Utility: assigning different costs to all four prediction results (see Figure 2).

Figure 2. Confusion matrix with cost values for utility evaluation.

	Patient deteriorates	Patient stable
EWS predicts deterioration	**True Positive (TP)** Flagged deteriorating (cost = **-20**, to reflect harm of recognized decline)	**False Positive (FP)** Incorrectly flagged (cost = **-1**, to reflect minor burden of manual checking and possible testing)
EWS predicts **no** deterioration	**False Negative (FN)** Missed deteriorating (cost = **-40**, to reflect harm associated with missed recognition of decline)	**True Negative (TN)** Correctly unflagged (cost = **0**, to reflect the cost of stable hospitalized patient)

Results

Study Population: Of 363,199 hospitalized patients, 16,826 (4.6%) experienced clinical deterioration.

Net Benefit:

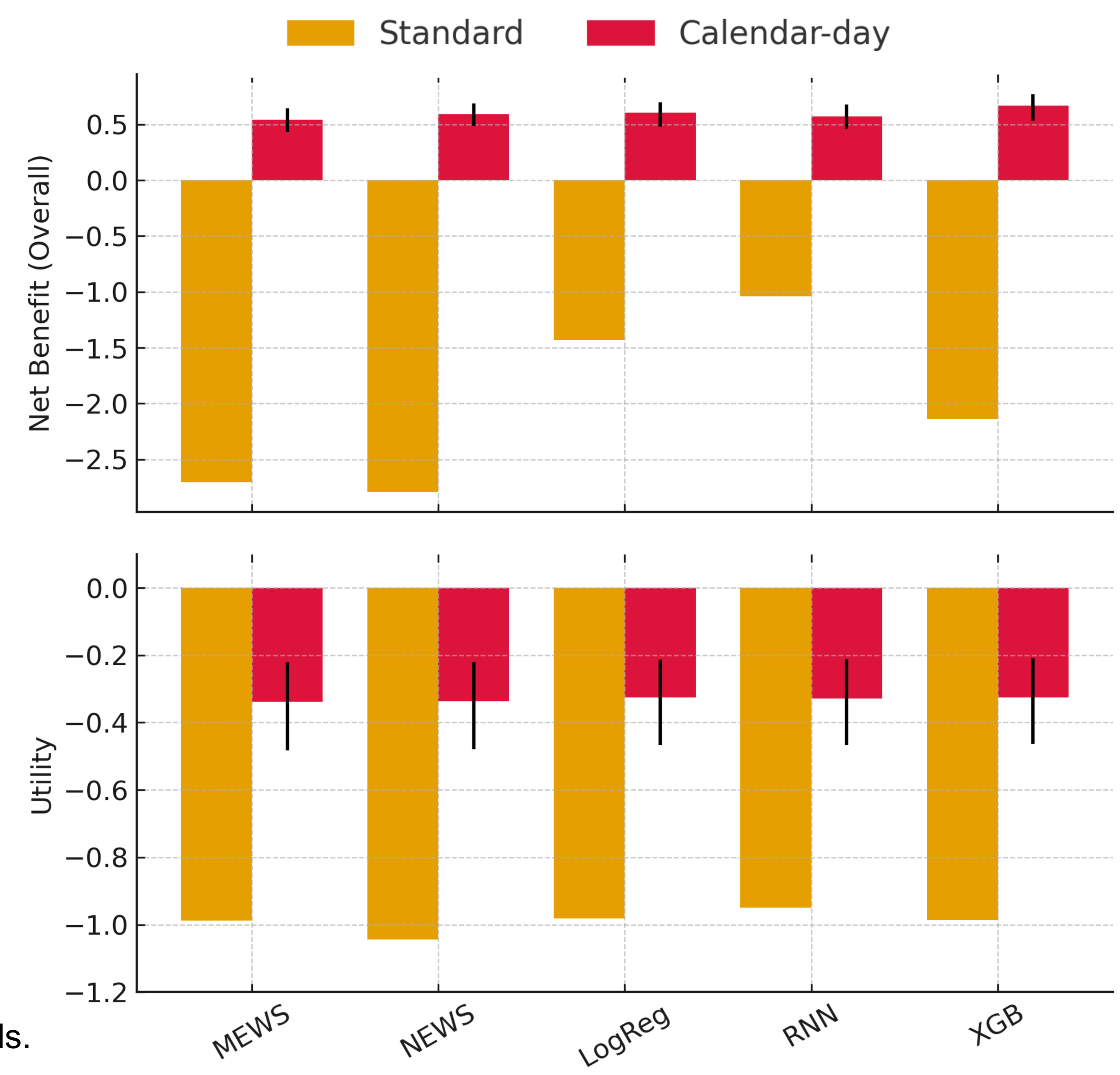
- Under the **standard pooling approach**, overall net benefits are low (–1.039 to –2.794).
- With the **calendar-day framework**, all EWS models showed **higher overall net benefit.**

Utility:

- Penalizing missed detection (FN) more heavily.
  - → Better reflects hospital priorities.
  - → Interpretable and clinically realistic.
- Calendar-day utility values of all EWS models were higher than their standard counterparts.

Best Performer: XGB achieved the highest calendar-day net benefit and utility, significantly outperforming other EWS models.

Figure 3. Comparison of standard and calendar-day net benefit and utility.



Conclusions

- Standard evaluation methods **pool across non-overlapping hospitalizations**, producing implausible or misleading net benefit values.
- Our **calendar-day framework** provides higher estimates of net benefit and clinical utility by evaluating patients hospitalized concurrently.
- **Clinical utility**, which allows for allocating greater harm to missed deterioration, better reflects hospital priorities than net benefit alone.
- **XGB** achieved the highest calendar-day net benefit and utility.
- Our framework facilitates better estimation of EWS impact during validation.

References

1. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM Int J Med*. 2001;94(10): 521-526. doi:10.1093/qjmed/94.10.521
2. McGinley A, Pearse RM. A national early warning score for acutely ill patients. *BMJ*. 2012;345:e5310. doi:10.1136/bmj.e5310
3. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Crit Care Med*. 2016;44(2):368-374. doi:10.1097/CCM.0000000000001571
4. Winslow CJ, Edelson DP, Churpek MM, et al. The Impact of a Machine Learning Early Warning Score on Hospital Mortality: A Multicenter Clinical Intervention Trial. *Crit Care Med*. 2022;50(9):1339-1347. doi:10.1097/CCM.0000000000005492
5. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128-38.

SCAN ME

Explore Clinical Informatics at ICU Data Science Lab!

Acknowledgements

Dr. Mayampurath, Ms. Wu, Mr. Carey, and Dr. Parker are supported by an R01 from NHLBI (R01 HL 173037). Ms. Strutz is supported by an NLM training grant (5 T15 LM007359).

Department of Medicine
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH