MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231202

Clinical Implementation of an AI Early Warning System Algorithm: Lessons Learned

Anne M. MEEHAN[a], Marcia A. CORE[b], Jared M. ROSS[b], Parvez A. RAHMAN[c], Bijan J. BORAH[c] and Pedro J. CARABALLO[a,d,1]
[a]*Department of Medicine, Mayo Clinic, Rochester, MN, USA*
[b]*Department of Information Technology, Mayo Clinic, Rochester, MN, USA*
[c]*Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA*
[d]*Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA*
ORCiD ID: Anne Meehan https://orcid.org/0000-0002-4296-5332,
Pedro Caraballo https://orcid.org/0000-0001-9050-7080

Abstract. The Deterioration Index (DI) is an automatic early warning system that utilizes a machine learning algorithm integrated into the electronic health record and was implemented to improve risk stratification of inpatients. Our pilot implementation showed superior diagnostic accuracy than standard care. A score >60 had a specificity of 88.5% and a sensitivity of 59.8% (PPV 0.1758, NPP 0.9817). However, acceptance in the clinical workflow was divided; nurses preferred standard care, while providers found it helpful.

Keywords. Early warning score, machine learning, artificial intelligence

1. Introduction

The primary goal of many hospital initiatives is identifying patients at risk of clinical deterioration. An ideal early warning system (EWS) tool should be automatic, accurate, and easy to use at the bedside. Structured big clinical data combined with machine learning (ML) and artificial intelligence (AI) offer the opportunity to achieve more accurate, real-time, and individualized predictions. Here we present preliminary results on the clinical implementation of an ML/AI algorithm-based early warning system.

2. Methods

The Deterioration Index (DI) (Epic, Verona, USA) is a commercially available, automatically calculated early warning algorithm integrated into the electronic health record that uses clinical and laboratory data to risk stratify patients. Patients are assigned a score out of 100 and defined as low (<30 green), intermediate (30-60 orange), or high risk (>60 red) of an adverse event (any cause mortality, cardiac arrest, transfer to intensive care, evaluation by the rapid response team). Retrospective validation was

[1]Corresponding Author: Pedro J. Caraballo, email: caraballo.pedro@mayo.edu.

performed on a sample of general-admission adult inpatients. DI performance was compared to our current institutional standard, the Modified Early Warning System (MEWS), a simple calculation using five bedside measures: SBP, HR, RR, temperature, and level of consciousness. Subsequently, a nursing-led pilot was pursued to assess the DI score's clinical utility on general inpatient admission floors. Nurses and providers (physicians and advanced practice providers) added the score to their workflow screen. Nurses were to alert providers to a new red score. The patient was then assessed for possible clinical deterioration. This workflow was similar to that already in use for the MEWS. Surveys were sent to collect perspectives on DI use, including a comparison to MEWS. DI scores were electronically collected for additional analyses during the same pilot period. T-test was used to compare the DI scores among patients who had an adverse event and those who did not.

3. Results

Retrospective validation of the DI score on general admission inpatients showed a C-statistic of 88.93% when a score of 50 was used as a cut-off for mortality (PPV 0.3547 and NPV 0.9936). This compared favorably to the MEWS score of 8, where the C-Statistic was 69.24% (PPV 0.1, NPV 0.993). During the pilot period, 2206 encounters were available for evaluation with a 3.94% prevalence of adverse events. A DI score >60 had a specificity of 88.5% and a sensitivity of 59.8% (PPV 0.1758, NPP 0.9817) for predicting an adverse event. When the distribution of the DI scores was compared between encounters with No-Event (n=2119) vs. encounters with Event (n=87), the DI scores 3 hours after admission (mean 31.96 (SD 10.92) vs. 42.92 (15.57)), the highest score during the encounter (43.55 (13.35) vs. 62.80 (17.98)) and the last score before the event or discharge (28.76 (8.42) vs. 54.34 (18.44)), were all statistically significantly different with p-values <.0001. Half of the adverse events occurred within 3 hours of the highest recorded DI score, and another 25% occurred in the following 39 hours.

Post-pilot survey responses were obtained from 47 providers and 44 nurses. When asked if the DI score helped with clinical care, 19% of the providers and 59% of the nurses said “No”. Approximately, 80% of providers reported that a DI score cutoff of 60 to escalate care was appropriate. Reasons that users disliked the DI score included lack of transparency as to how the score is calculated, persistently high scores without apparent deterioration due to chronic conditions, low scores in patients with apparent deterioration, and the wide time range between an increase in the score and occurrence of an adverse event.

4. Conclusions

The DI algorithm performed well in detecting patients at risk. A DI score of >60 has an excellent NPV of 99%, but the PPV and sensitivity are low when the prevalence of events is low. Providers found the score helpful and a cutoff of sixty appropriate as a signal of deterioration. On the other hand, nurses preferred to use the MEWS to direct patient care. Age and chronic comorbidities may contribute to persistently high DI scores without acute deterioration, which limits clinical value. The black box phenomenon of AI predictive models cannot be underestimated at the time of clinical implementation. Good statistical accuracy does not guarantee clinical acceptance.