Evaluating Deterioration Prediction, Usability, and Impact of a Clinical Artificial Intelligence System: Real-time Intensive Care Warning INdex System (I-WIN) Using EHR and Bedside Monitor Data

Fuchiang (Rich) Tsui, PhD[1-3], Victor Ruiz, PhD[1], Sachin Grover, MS[1], Lingyun Shi, MS[1], Allan Simpao, MD[2,3], and Michael Goldsmith, MD[2,3]

[1]Tsui Laboratory, [2]Department of Anesthesiology and Critical Care Medicine, Children's Hospital of Philadelphia; [3]School of Medicine, University of Pennsylvania, PA

Introduction: Intensive care units (ICUs) face significant challenges despite continuous investment in medical technology and personnel. These include high mortality rates compared with other hospital units, increased admissions, medical errors, staff shortages, and alert fatigue.[1] Infants with single-ventricle physiology are a complex subpopulation with elevated mortality and morbidity risk before stage-2 palliation.[2] Early detection of deteriorating physiology in these patients may help decrease critical decompensation events and allow timely interventions. Despite advances in artificial intelligence (AI) in clinical decision support,[3] several challenges remain, e.g., lack of personalized modeling due to the limited number of expert-chosen variables from the electronic health record (EHR), and a lack of trust in "black box" models which decreases adoption even with improved accuracy. This study aimed to conduct a comprehensive, 3-stage evaluation of the Intensive Care Warning INdex System (I-WIN) system. I-WIN includes displays of streaming bedside vital-sign waveform data, a model for deterioration prediction with personalized explanations, and a multi-modal data repository. We hypothesize I-WIN achieves clinical impact by improving clinicians' ability to forecast severe deterioration in ICU patients.

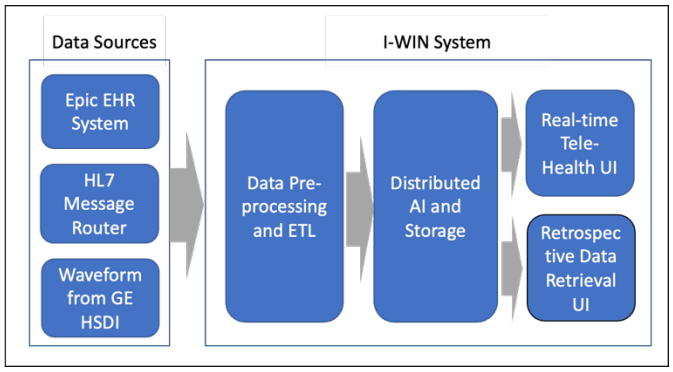

Figure 1. Architecture of the Intensive care Warning Index (I-WIN) system. HSDI: high speed data interface; UI: user interface.

Methods: The CHOP Institutional Review Board (IRB) approved this study.

I-WIN System Description: I-WIN provides a real-time view of patients' vital-sign waveforms along with clinical EHR data. It stores historical data within a data repository and provides early and accurate prediction with explanation of deterioration events including extracorporeal membrane oxygenation (ECMO) cannulation, cardiac arrest, and emergent endotracheal intubation (EEI). **Figure 1** shows four key process layers in I-WIN: data sources, an extract-transform-load (ETL) layer, distributed AI, information presentation, and data warehouse. Our deployment follows industrial standards: security using organizational user authentication and a secure interface, system deployment using Docker and Kubernetes, distributed computing, and user accountability.

Cohort Data for Predictive Modeling The model evaluation cohort included infants with palliated or unrepaired single-ventricle, ductal-dependent, or shunt-dependent congenital heart diseases admitted to a cardiac intensive care unit in an urban academic children's hospital and is further described in our previous work.[4] Cases were defined as a composite of critical deterioration events: emergent endotracheal intubation, cardiopulmonary resuscitation (CPR), extracorporeal membrane oxygenation cannulation, or CPR with refractory cardiac arrest requiring ECMO.

Stage I: Deterioration Prediction Evaluation Stage I was a retrospective evaluation of deterioration predictions in the cohort described above. Models were built from 1,028 variables and featured an ensemble of extreme gradient boosting (XGB) classifiers.[4] Evaluation metrics included area under the ROC curve (AUC), sensitivity, specificity, and positive predictive value at five prediction horizons (1, 2, 4, 6, 8 hours before the onset deterioration events).

Stage II: Usability and Workflow Evaluation Stage II was an evaluation of I-WIN's usability and workflow through a web-based interactive demonstration system featuring real-time vital signs and waveforms, vital-sign trends, deterioration risk estimates, and personalized risk explanations. Each usability-test user used the demonstration system and completed a usability and workflow survey. The survey had five categories: 1) user profile, 2) needs assessment,

3) system usability scale (SUS) and workflow, 4) post-deployment open-ended questions, and 5) future system use. The survey was deployed in a cloud-based survey tool independent of I-WIN.

Stage III: Clinical Impact Evaluation Stage III consisted of a clinical-impact survey form to measure 1) clinicians' ability to predict future patient deterioration, and 2) the efficiency (time needed) of clinicians when assessing patients' status in real clinical scenarios. This evaluation had two arms, namely a decision-support (intervention) arm and a treatment-as-usual (TAU) arm with an equal number of patients to be reviewed by clinical impact-test users (clinicians). The intervention arm provided a summary of EHR data along with I-WIN risk predictions and explanations. The TAU arm presented only the summary EHR data. The summary EHR data included screenshots of the RN ICU summary page, which shows a time series of medications, and vital signs from the Epic EHR system used by clinicians in their daily practice. In both arms, 50% of patients underwent CPR. Each test user was randomly assigned 12 patients from a subset of single-ventricle cohort in two arms (6 for each arm), i.e., each arm comprised 3 patients with cardiac arrests and 3 patients without any events. Evaluation metrics included sensitivity, specificity, accuracy, and response time.

Results: Stage I: Deterioration Prediction Evaluation The details of prediction evaluation using multiple machine learning models were previously published.[4] Specifically, at 4 hours before deterioration, the XGB model achieved an area under the receiver operating characteristic curve (AUC) of 0.92 (95%CI: 0.84-0.98), 0.88 sensitivity, 0.78 positive predictive value, 0.86 specificity, and 0.57 Brier skill score.[4]

Stage II: Usability and Workflow Evaluation Ten users participated in the usability and workflow evaluation. Users voluntarily responded to broadcast emails from authors (MG, AS) with links to the I-WIN demonstration system and a user survey. 50% (n=5) of respondents reported experience using early warning systems besides EHR systems. In the needs assessment category, all users agreed with the statement, "Having summary information from multiple data sources saves time and improves workflow". The average SUS score was 73.5, representing an acceptable system.

Stage III: Clinical Impact Evaluation Thirteen CICU clinicians participated in the evaluation. Each test user reviewed 12 patients randomly selected from a pool of 48 patients (24 [50%] with cardiac arrest and 24 [50%] without any critical events) following stratified sampling and resampling approach. A total of 156 survey patients were reviewed by the 13 test users. **Table 1** shows the accuracy, sensitivity, specificity, and average process time per patient between the two arms. Compared to the users in the TAU arm, users in the intervention arm showed improved sensitivity, specificity, accuracy, and information process time by 32.2%, 9.6%, 18.5%, and 27.1%, respectively.

Table 1: Stage III performance comparison between an intervention and the treatment-as-usual arms.

Survey patients (n=108); Test Participants (n=9)	Average Sensitivity (std; P-value)*	Average Specificity (std; P-value)	Average Accuracy (std; P-value)	Average Process Time per Patient in seconds (std; P-value)
Intervention Arm [with I-WIN] (n=78)	**0.74 (0.24; 0.17)**	**0.87 (0.22; 0.39)**	**0.81 (0.18; 0.12)**	**86.99 (56.2; 0.44)**
Treatment-as-usual Arm [without I-WIN] (n=78)	0.56 (0.29; ref)	0.8 (0.22; ref)	0.68 (0.17; ref)	119.32 (185.52; ref)

*: std: standard deviation; all P-values were from two-sided paired t-tests.

Discussion We successfully tested that the IWIN system has a positive impact on deterioration prediction. Most users expressed the need for aggregate summary information from multiple data sources and a high burden of false alarms in the ICU in alignment with evidence. In stage III, we found that the intervention arm (with I-WIN support) showed improved sensitivity, specificity, accuracy, and timeliness of cardiac arrest recognition compared to the TAU arm. Our study has limitations. This is a single-center study conducted in a quaternary care children's hospital. Limited numbers (Stages 2 and 3: n=10 and 13, respectively) of test users participated in the 2nd and 3rd stages of evaluation.

Conclusions We deployed a real-time data-driven AI system I-WIN[3,4] in a busy cardiac intensive care unit at an urban quaternary care children's hospital, which not only collects and stores streaming waveform and EHR data, but also predicts deterioration risk with explanations. With the adoption of industry standards for deployment, I-WIN can be potentially deployed in different hospital settings.

References

1. Society of Critical Care Medicine. Critical care statistics.
2. Tabbutt S, et al. Risk factors for hospital morbidity and mortality after the Norwood procedure: A report from the Pediatric Heart Network Single Ventricle Reconstruction trial. *J Thorac Cardiovasc Surg*. 2012;144(4):882-895.
3. Ruiz VM, Saenz L, Lopez-Magallon A, Shields A, Ogoe HA, Suresh S, Munoz R, Tsui FR. Early prediction of critical events for infants with single-ventricle physiology in critical care using routinely collected data. *J Thorac Cardiovasc Surg*. 2019;158(1):234-243.
4. Ruiz V, Goldsmith M, Shi L, Simpao A, Galvez J, MY N, Nadkarn V, Gaynor W, Tsui F. Early prediction of clinical deterioration using data-driven machine learning modeling of electronic health records. *J Thorac Cardiovasc Surg*. July 2022.