Title

Generalizability of SuperAlarm via Cross-Institutional Performance Evaluation

Authors

Xiao, Ran
Do, Duc
Ding, Cheng
et al.

Publication Date

2020

DOI

10.1109/access.2020.3009667

HHS Public Access

Author manuscript

IEEE Access. Author manuscript; available in PMC 2021 March 18.

Published in final edited form as:

IEEE Access. 2020 ; 8: 132404–132412. doi:10.1109/access.2020.3009667.

Generalizability of SuperAlarm via Cross-Institutional Performance Evaluation

Ran Xiao[1,2] **[Member, IEEE]**, **Duc Do**[3], **Cheng Ding**[1], **Karl Meisel**[4], **Randall Lee**[4], **Xiao Hu**[1,2] **[Senior Member, IEEE]**

[1]School of Nursing, University of California San Francisco, San Francisco, CA 94143 USA

[2]School of Nursing, Duke University, Durham, NC 27708 USA

[3]UCLA Cardiac Arrhythmia Center, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095 USA

[4]School of Medicine, University of California San Francisco, San Francisco, CA 94143 USA

Abstract

Bedside patient monitors are ubiquitous tools in modern critical care units to provide timely patient status. However, current systems suffer from high volume of false alarms leading to alarm fatigue, one of top technical hazards in clinical settings. Many studies are racing to develop improved algorithms towards precision patient monitoring, while little has been done to investigate the aspect of algorithm generalizability across different health institutions. Our group has been developing an evolving framework termed SuperAlarm that extracts multivariate patterns in data streams (monitor alarms, electronic health records and physiologic waveforms) of modern health enterprise to predict patient deterioration and has demonstrated great potential in mitigating alarm fatigue. In this study, we further investigate the generalizability of SuperAlarm by designing a comprehensive approach to achieve performance comparison in predicting in-hospital code blue (CB) events across two health institutions. SuperAlarm model trained with alarm data in one institution is tested on both internal and external test sets. Results show comparable performance with sensitivity up to 80% within one-hour window of events and over 90% in reduction of false alarms in both institutions. Cross-institutional performance agreement can be further improved by predicting a more stringent CB subtype (cardiopulmonary arrest), with internal sensitivity lying within 95% confident interval of external one up to 8-hour before event onset. The cross-institutional performance comparison offers first-hand knowledge on both advantages and challenges in generalizing a prediction algorithm across different institutions, which hold key information to guide the design of model training and deployment strategy.

Keywords

Patient monitoring; predictive models; biomedical informatics

Corresponding author: Ran Xiao (ran.xiao@duke.edu).

I. INTRODUCTION

Bedside patient monitors are essential clinical devices in acute care settings that provide timely information about patients' physiologic condition. However, current patient monitors are known to produce excessive alarms with the majority of them either false or non-actionable [1]-[6]. A variety of factors individually or collectively contribute to the excessive false alarms including inappropriate system settings, suboptimal signal quality and deficiencies in proprietary algorithms [7]. Another significant contributor to the excessive alarms is nuisance alarms. Nuisance alarms are usually non-actionable and triggered by transient perturbation in patients' status crossing the threshold that trigger on/off alarms, but do not necessarily carry much information indicative of major health status change [8]-[10]. Bedside caregivers can develop alarm fatigue as they are constantly exposed to visual and auditory sensory overload from excessive alarms in a typical 8–12 hour shift [11]-[14]. This creates an unsafe clinical environment as alarms of impending adverse events might be overlooked among false or nuisance ones, resulting in delayed or even missed opportunity for timely intervention [15]-[18]. This may also inflate the stress level of patients as these alarms direct act at the bedside leading to a decline in the quality of patient care.

A surge in effort is being made towards a more precise patient monitoring solution by improving detection algorithms of specific types of alarms aiming to reduce false alarms. Many studies direct analyze physiologic signals (electrocardiography (ECG), photoplethysmography (PPG), arterial blood pressure waveform, etc.) to improve detection accuracy of cardiac arrhythmia alarms in intensive care [19]-[22]. There are also a collection of studies using data mining and machine learning techniques to analyze information in electronic health record (EHR) systems to derive early warning scores that can provide early warning to clinicians about patient deterioration [23]-[26]. Our group has been developing an evolving framework, the SuperAlarm, that extracts hidden multivariate patterns in multi-modality data streams (monitor alarms, electronic health records and physiologic waveforms) of modern health enterprise as features to predict a target clinical endpoint [27]-[30].

In contrast to other methods, SuperAlarm provides a unique strategy that directly acts upon monitor alarms while remaining flexible to incorporate other data modalities. A series of studies have been conducted that delineate a roadmap for the development of SuperAlarm framework through predicting in-hospital code blue events (i.e., cardiopulmonary arrest (CPA), acute respiratory compromise (ARC) and other medical emergencies (Others)). First, we extracted frequent co-occurring patterns (termed SuperAlarm patterns) based on the Apriori algorithm[31] using monitor alarms preceding code blue events[29]. Next we focused on expanding the SuperAlarm framework by incorporating other data modalities that are readily available in a connected healthcare enterprise to further improve performance [30]. The next two studies further extended SuperAlarm framework by integrating cumulative effect and temporality in the sequences of SuperAlarm patterns into model training[27], [28].

The present study aims to further investigate the SuperAlarm framework with regard to its generalizability. On the one hand, well generalized performance of SuperAlarm model

indicates potential knowledge transfer (such as pretrained model parameters) from one institution to another without the need for model training from scratch. On the other hand, for rare clinical end points, well generalized performance of SuperAlarm model opens up the potential of enlarging training samples by combining data from multiple institutions towards training more complex and precise models. The present study adopted consistent framework as our previous studies to derive SuperAlarm patterns with integration of their temporal relationships as features to train the final prediction model [28], [30], while offering novel insights into its generalizability by laying out a comprehensive approach to assess the internal and external performance in predicting in-hospital code blue events. To this end, a data preprocessing pipeline was designed to harmonize monitor alarm data from two health institutions, and the predictive model was trained solely with data from one institution so that its cross-institutional performance could be assessed. The obtained results shed light on the generalizability of SuperAlarm, as well as the advantages and challenges in generalizing SuperAlarm across different institutions.

II. METHODS

A. DATA DESCRIPTION

Alarm data from bedside patient monitors located in various intensive care units (ICUs, including neurosurgical, cardiothoracic, coronary care, medical units, etc.) and other critical care units (cardiac observation, hematology and stem cell transplant, medical-surgical, neuroscience and stroke units, etc.) were obtained from two healthcare institutions, University of California San Francisco (UCSF) Helen Diller Medical Center and University of California Los Angeles (UCLA) Ronald Regan Medical Center. Bedside patient monitors from which alarm data was extracted were from the same vendor (GE Healthcare, Milwaukee, WI) in both institutions. Hospital encounters with documented code blue events were selected as cases. For patients with multiple code blue events, only encounters with the first code blue event were selected for analysis. Controls consisted of hospital encounters of matching time period as those in case condition but without documented code blue events and unplanned ICU transfer. The control encounters were then subject to the following screening criteria: (1) match the same all patient refined diagnosis-related group (APR-DRG) or Medicare DRG; (2) the same gender; (3) within ±5 years of age; (4) in the same medical units as case encounters. There were 412 code blue encounters (2,099,026 alarms) and 4020 control encounters (12,696,925 alarms) between 2013 and 2018 selected from UCSF with mean age 60.8 ± 16.1 and 58.9% male. There were 254 code blue encounters (662,576 alarms) and 2213 matched control (5,363,019 alarms) encounters between 2010 and 2012 selected from UCLA with mean age 61.3 ± 17.9 and 54.2% male. The Institutional Review Board (IRB) from both institutions approved the analysis of patient data with a waiver of patient consent.

B. ALARM DATA PREPROCESSING

Alarm data could vary across institutions in numerous ways, such as wording of alarm messages, institutional guideline for setting alarm thresholds, types and frequencies of alarms. A comprehensive alarm preprocessing procedure was designed to harmonize alarm data before model training. First step was to make alarm message port agnostic. Same types

of alarms could be collected from any port on the device, which was inscribed in alarm messages and needed to be removed. For instance, the alarm message “ART1 S HI” (high systolic arterial blood pressure alarm from port 1) was updated to “ART S HI” to remove the port information. Next, alarms that measure same physiological status via noninvasive approach and its invasive counterparts were merged and treated equally, e.g. blood pressure alarms generated from noninvasive calf and invasive arterial measurement. Finally, alarms from both institutions were mapped into a common set of alarm codes to fix any differences in the wording of alarm messages, e.g. “BP DIA HI” at UCSF vs. “BP D HI” from UCLA were all mapped to an alarm code “21”, which denotes the “high diastolic blood pressure” alarm. The alarm mapping table with a full list of alarm codes can be found in the Supplement.

With mapped alarm codes, additional preprocessing steps were performed. First, alarm codes related to technical alarms (e.g., “ECG LEADS FAIL”) that didn’t add predictive value to the model training were removed. Second, system-defined crisis alarms (i.e., “ASYSTOLE”, “VFIB/VTAC” and “APNEA”) which are clear indicators of emergency situation and usually take place close to code blue event onset were removed to avoid artificially boosting the predictive power of SuperAlarm. Third, parametric alarms with continuous numeric measurement, such as alarms associated with blood pressure, heart rate and SPO2, were discretized by a class-attribute contingency coefficient (CACC) based discretization algorithm[32], which offers a discretization scheme with considerations of interdependence between class and discretized attribute to facilitate subsequent model learning. Finally, case encounters that contained unusually low number of alarms likely due to technical issues were removed. In brief, this was achieved by modeling the arrival of alarms with a Non-Homogenous Poisson Process (NHPP) for all case encounters to derive the 95% confidence interval (CI) of the mean alarm count [33]. The lower bound of CI then served as a threshold for minimal alarm count to be considered as training encounters.

C. SUPERALARM PREDICTIVE MODELING

With preprocessed alarms, the SuperAlarm model was developed following the procedure illustrated in Fig. 1. The current study followed the same framework as our previous studies [27], [28] hence is briefly described here. As illustrated in the left panel of Fig. 1, selected time windows (T_w) of alarms from cases and controls were used to mine SuperAlarm patterns with the frequent itemset mining algorithm, Maximal Frequent Itemset Algorithm (MAFIA) [34], [35]. Conceptually, the patterns fulfill the following criteria, they frequently occur in cases surpassing a minimal support (SUP_{min}) while seldomly occur in controls with a frequency below a predefined false positive rate (FPR_{max}). Different time windows ([0.5, 1, 1.5, 2 hours]) and minimal support values ([0.1, 0.15, 0.2]) were explored during the model training process. Optimal values of T_w and SUP_{min} were determined through hyperparameter tuning process under various preset FPR_{max} ([0.1, 0.15, 0.2, 0.25]), resulting in four final sets of SuperAlarm patterns. Each alarm in time were then transformed into a binary vector designating whether each SuperAlarm pattern was triggered in the preceding Tw window of the alarm.

The sequences of SuperAlarm pattern triggers, or SuperAlarm trigger sequences, then served as samples to train the prediction model as shown in the right panel of Fig.1. SuperAlarm trigger sequences were sampled differently in case and control conditions. 5 trigger sequences were selected from each control encounter by randomly picking an anchor time point as the ending point of a sequence (i.e., sequence anchor) based on a uniform distribution across the whole encounter timeline. Each sequence was then constructed by collecting all triggers between first trigger in the encounter and the anchor trigger. For case encounters, the number of trigger sequences to be selected was 5 multiplied by the ratio of encounter count between control and case conditions to achieve a balanced training set. The selection of case trigger sequences was based on an exponential distribution with higher probability of sequence anchors being sampled near the code blue onset under the heuristics that more information about code blue events can be captured in this way. The profile of exponential distribution was determined by

$$p(t) = \mu \cdot e^{-\frac{t}{\mu}} \tag{1}$$

where t is time before code blue onset; $P(t)$ is the probability of trigger at time t being selected as the sequence anchor; μ is the mean of exponential distribution, which was set as [10, 30, 60 minutes] as a hyperparameter to tune. To account for the cumulative effect and temporality of trigger sequences, the weighted average occurrence representation (WAOR) method[27] was adopted to assign higher weights for triggers close to the sequence anchor and take the weighted sum of all triggers in the sequence, by

$$WAOR(m) = \sum_{t=1:t_{an}} 1/\left(\left(\frac{|t_{an}-t|}{\gamma}\right)^{a} + \beta\right) \cdot T(t,m) \tag{2}$$

where $1/\left(\left(\frac{|t_{an}-t|}{\gamma}\right)^{\alpha} + \beta\right)$ is the WAOR weighting function that assigns weight to trigger $T(t, m)$ at time t and SuperAlarm pattern m based on its temporal distance to the anchor trigger, i.e. $|t_{an} - t|$. The temporal profile of the weighting function was controlled by three parameters, α ([0.5, 1, 2, 3]), β ([0.1, 0.5, 1]) and γ ([0.5, 1, 10, 100]). These parameter candidates were chosen based on an extension of parameters from our previous study that demonstrated the feasibility of WAOR in capturing temporal dynamics of trigger sequences [27]. The expanded choices covering a much larger range aim to derive optimal WAOR weighting function through a more systematic way. The derived WAOR weights from each trigger sequence were rescaled by min-max normalization across all training samples and then served as input features to train the final prediction model based on a logistic regression classifier with lasso regularization. The regularization parameter λ was a hyperparameter to tune and was selected from a logarithmically spaced vector of 50 numbers between 10^{-3} and 10^{-1}.

All hyperparameters, including one sampling parameter during trigger sequence selection, three parameters during WAOR weighting process and one regularization parameter during classifier training, were tuned simultaneously through 5-fold cross validation using the training set. The optimal combination of these hyperparameters was determined by the one

with the highest average area under the receiver operating characteristics (AUC), based on which a final prediction model was derived using the whole training set.

D. CROSS-INSTITUTIONAL PERFORMANCE EVALUATION

To test the generalizability of SuperAlarm, training and test sets from two institutions were selected as following. Alarm data from UCSF were divided into training and test sets at 80% −20% split at the encounter level. The UCSF training set was used to develop the SuperAlarm model, which was first evaluated for its internal performance on the UCSF test set. Internal performance was evaluated in both offline and simulated online fashions. During offline test, SuperAlarm trigger sequences were sampled in the test set the same way as the training set. Conventional metrics, including AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and F1 score, were calculated to evaluate the offline performance.

During simulated online test, the model was evaluated continuously for each alarm in time in testing encounters to simulate the real-life scenario during model implementation at bedside patient monitors. During this setup, each alarm and its preceding Tw window of alarms were accessed to generate a binary vector of SuperAlarm pattern triggers representing whether each SuperAlarm pattern was matched. Then WAOR-based aggregation of each trigger vector joint with all of its historic triggers formed a feature vector, with which the trained logistic regression model made prediction for the risk of code blue events. To tackle the continuous nature of simulated online tests, three time-aware metrics were adopted to evaluate the model performance. Sensitivity along lead time (Sen@T_L) was the proportion of case encounters with at least one prediction of y = 1 (predicted as code blue) within 12-hour time window preceding T_L (lead time before code blue onset) in the test set. Alarm frequency reduction rate (AFRR) measured the average hourly reduction in alarm frequency contributed by SuperAlarm algorithm comparing to regular monitor alarms. Work-up to detection ratio (WDR) resembles number needed to treat (NNT), and was calculated as following

$$WDR(prev) = \frac{TPR * prev + FPR * (1 - prev)}{TPR * prev} \quad (3)$$

where TPR(true positive rate) was the ratio of correctly predicted case encounters over total case encounters in the test set within 12-hour window preceding code blue onset; FPR(false positive rate) was the average ratio of falsely predicted control encounters (y = 1) over total control encounters in the test set across 1000 randomly selected 12-hour window in each control encounter. Importantly, the calculation of WDR took account of prevalence ($prev$) of code blue events, a critical component in designing unbiased performance metrics within the clinical context [36].

Following the same simulated online fashion, the model was tested externally on UCLA dataset in the following two ways. The whole UCLA dataset first served as one test set to evaluate the overall external performance, which was qualitatively compared to the internal performance. Alternatively, 100 testing sets (each with the same number of samples as in UCSF test set) were extracted from the UCLA dataset via bootstrapping so that a

distribution of external performance could be estimated[37]. Cross-institutional performance was quantitatively compared by evaluating the internal Sen@T_L curve against 95% confidence interval (CI) of external Sen@T_L. In addition, internal and external testing AFRR and WDR were compared by one-sample Student's t-test with Bonferroni correction for multiple comparisons (α = 0.025) [38].

III. RESULTS

A. SUPERALARM MODELS

The number of SuperAlarm patterns ranged from 377 at FPRmax of 0.1 to 798 at FPRmax of 0.25, as shown in Table I. It also lists a full list of hyperparameters deriving SuperAlarm patterns. Table II lists the most frequent patterns at each cardinality (number of alarms in a SuperAlarm pattern) 12 hours preceding code blue events based on FPRmax of 0.25. It shows that SuperAlarm patterns mainly consisted of alarms reflecting cardiorespiratory status (e.g., blood pressure) and ECG arrythmia alarms. For instance, SuperAlarm pattern ["BP RATE LO: [,33.5]", "BP SYS LO: [70.5, 82.5]", "PVC"] include three alarms that co-occur in a window of 1.5 hours, a low pulse rate alarm (derived from arterial blood pressure) lower than 33.5 beats/minute, a low systolic blood pressure alarm between 70.5 and 82.5 mmHg and a premature ventricular contraction (PVC) arrythmia alarm.

Fig. 2(a) illustrates various WAOR profiles being explored as controlled by (2) to vectorize SuperAlarm trigger. By design, they all presented a monotonic trend with larger weights towards code blue events, with the green curve as the optimal profile selected during the training process. Fig. 2(b) shows WAOR weights for all case sequences and matched number of control sequences in the training set. It shows more prominent weights across most SuperAlarm patterns in case sequences on the left than those from control on the right.

B. INTERNAL AND EXTERNAL PERFORMANCE OF SuperAlarm

Internal performance of SuperAlarm through offline evaluation is presented in Table III It shows highest AUC at 85.96% was achieved by model trained with FPRmax at 0.25. The model achieved 81.13% of sensitivity while maintaining a specificity of 78.89%. The model achieved the best NPV at 93.33% while delivered the second best PPV at 53.46%. Fig. 3 (a)–(c) present the internal performance of SuperAlarm through simulated online evaluation. Predictive power of SuperAlarm increased as prediction window closer to code blue events with sensitivity at 50%~80% depending on the selection of FPR_{max} (see Fig. 3(a)). It shows best performance was achieved with the model based on FPR_{max} of 0.25, with sensitivity in the range of 70%~80% within one-hour window of event onset. Fig. 3(b) shows WDR fell in the range of 5–8 at current prevalence of 0.055, which also presents its changing trend with respect to the prevalence of code blue events. Nonetheless all models achieved hourly reduction in false alarm rate (AFRR) at over 90% (see Fig. 3(c)).

Fig. 3(e)–(f) show the external performance of SuperAlarm on the UCLA dataset. Similar to internal performance, it showed external sensitivity fell in the range of 50%~80% preceding code blue events and degenerated with increasing lead time (see Fig. 3(a)&(e)). However, they presented larger range of WAORs between 5–9, which went up to 6–10 if matching the

prevalence to internal test set (see Fig. 3(b)&(f)). Nonetheless, external performance offered slightly better AFRR at 95%~98% compared to the internal one at 90%~97% (see Fig. 3(c)&(g)).

C. STATISTICAL COMPARISON OF CROSS-INSTITUTIONAL PERFORMANCE

Fig. 4(a) compares sensitivity across lead time between internal and external test sets. It shows internal sensitivity exceeded external one across majority of lead time. When comparing other two metrics, external performance achieves higher AFRR at 94.9% than internal AFRR at 93.1% ($p < 0.025$) at a cost of a higher WDR at 10.3 against internal one at 7.1 ($p < 0.025$).

As code blue events consist of a collection of medical emergencies with heterogeneous etiologies, incidence ratio of different subtypes was evaluated to probe the discrepancy of cross-institutional performance. Although CPA demonstrated to be the major subtype for both institutions (63.8% at internal dataset and 73.1% in external dataset), other two subtypes presented distinct proportions between two institutions, as shown in Fig. 5.

A subsequent comparison of performance was performed by keeping the same dominant subtype (i.e., CPA) in both test sets. Fig. 4(b) reveals conforming sensitivity between internal and external performance with matched code blue events, with internal sensitivity falling within 95% CI of external sensitivity across majority of time before event onset.

IV. DISCUSSION

The present study further examines the SuperAlarm framework with regard to cross-institutional generalizability aspect of SuperAlarm, which has not been investigated yet carries great implications to model design and deployment of SuperAlarm. To this end, the present study follows the consistent SuperAlarm framework as our previous studies [27], [28] but with much larger cross-institutional datasets and selection of model parameters, while focuses on developing a comprehensive scheme to test the generalizability of SuperAlarm across different institutions. A systematic approach has been designed from alarm curation, to training-test data arrangement, and all the way to performance evaluation to test the internal and external performance of SuperAlarm. The results show similar yet subtle differences in performance patterns between internal and external testing. The cross-institutional performance comparison offers first-hand knowledge that sheds light on both advantages and challenges in generalizing SuperAlarm across different institutions.

A. GENERALIZABILITY OF SUPERALARM

SuperAlarm demonstrates similar changing trends in all performance metrics when tested on internal and external test sets, as shown in Fig. 3. Both internal and external performance demonstrates comparable sensitivity close to code blue onset in the range of 50%~80%, which degenerates along with increase of lead time. Meanwhile, both internal and external performance shows significant reduction in alarm frequency with AFRR over 90% given any selected FPR_{max} thresholds. The consistency in performance is in part contributed by inherited features of the SuperAlarm framework that are in favor of generalizing the algorithm across institutions. First, the preprocessing steps harmonize alarm data to be

agnostic to device ports. Device port numbers are captured in alarm messages but are obviously not relevant to the objective of predicting patient status. Second, the SuperAlarm framework taking a tokenizing approach for any input information not only normalizes various data modalities but also offers a common code mapping scheme for data naming discrepancies across institutions - this approach allows us to apply domain knowledge to developing a common set of terms to describe alarms from different sources. Third, the discretization step for processing parametric alarms takes account of numeric measurements of these alarms[32], which helps mitigate policy discrepancies across institutions on threshold settings that trigger monitor alarms. These features are particularly effective for alarms but their effectives at processing other data modalities such as those from electronic health record systems within the SuperAlarm framework remain to be studied.

Despite promising similarity in performance, there are also challenges facing the generalization of SuperAlarm across institutions. The disparity in subtype distributions of code blue events (see Fig. 5) in two institutions plays a substantial role in performance discrepancy (see Fig. 4(a)). It can be seen that the discrepancy in performance can be largely reduced when predicting the same dominate subtype (see Fig. 4(b)). Such an improvement in agreement of sensitivity indicates generalization of SuperAlarm benefits from a clear and specific target clinical endpoint. Otherwise, generalizability of SuperAlarm might be discounted by different composition of clinical events in a different institution. Although at similar level, there also exist cross-institutional differences in AFRR and WDR ($p < 0.025$). One plausible cause is the discrepancy in alarm counts driven by different institutional policies on the setup of bedside patient monitors that impact the type and number of alarms generated. The two institutions indeed present clear differences in alarm counts, with average number of alarms at 4985 per case encounter and 3111 per control encounter at UCSF whereas those numbers are 2157 and 1922 at UCLA. Therefore, the evaluation of performance metrics needs to take these contexts into account.

B. STRENGTHS AND LIMITATIONS

A growing body of studies have been carried out to pursue precision patient monitoring and to battle alarm fatigue in critical care settings. Several studies aim to improve physiologic signal processing and detection algorithms for individual alarms [19]-[22], [39]-[42], whereas others utilize EHR information to derive new early warning metrics [23]-[26], [43]-[47]. SuperAlarm provides a unique strategy of directly analyzing alarm data. In addition, the SuperAlarm framework by design is flexible enough to take advantage of multiple data modalities available in the data stream in most model hospital settings to improve both sensitivity and specificity instead of relying on single data source [30].

Compared to single monitor alarms, SuperAlarm offers many advantages evidenced by both current and previous studies [27]-[30], [48], [49]. SuperAlarm takes a multivariate approach to extract co-occurring alarms within a time window and integrate their high-order interactions. This feature helps reduce false alarms since certain cooccurring alarms in SuperAlarm patterns provide mutual support that improves alarm fidelity. For example, a SuperAlarm pattern [“BRADY”, “BP DIA LO”] with arrythmia alarm “bradycardia” co-occurring with “low diastolic blood pressure” alarm that reflects the underlying

hemodynamic state is more likely to be true than an isolated “bradycardia” alarm. The later can falsely arise from poor ECG signal quality or deficiencies in proprietary arrythmia detection algorithms. The multivariate nature also enables the capture of physiological changing trend whereas single alarms can only offer one static status in time. The SuperAlarm framework also improves predictivity of impending clinical adverse events by capturing temporal dynamics over a period of time. In addition, SuperAlarm model is interpretable, which provides understandable patterns (see Table II) by clinicians in support of further diagnostic work-ups.

In addition to the level of agreement in performance between internal and external institutions, SuperAlarm performance achieved in the present study is consistent with our previous study that predicts in-hospital code blue events with monitor alarm data[29]. Both offer sensitivity in the range of 50~80% within 1-hour of event onset while maintaining over 90% AFRR. While unbiased performance comparison of SuperAlarm to other studies is challenging due to discrepancies in selection of target clinical events, data modalities, performance evaluation protocols, etc., best AUC (85.96%) achieved by SuperAlarm is on par with other machine learning based models for predicting cardiac arrest using vital signs (best AUC at 78.1%, 85% and 88.6% respectively), all of which outperform modified early warning score (MEWS, with AUC generally below 70%) [50]-[52].

At the moment, our data sources are limited to monitor alarm data, which are the common data modality available to us from both institutions. Nonetheless, being the backbone in the SuperAlarm framework, evaluating the generalization of SuperAlarm with monitor alarms still carries great significance as the starting point. Future effort will be directed towards collecting additional data modalities across institutions which can provide further insight regarding the generalizability of SuperAlarm. It will also enable objective comparison of SuperAlarm with other methods that rely on a data modality other than monitor alarms. It is also worth noting that the present study uses a simple mapping scheme to consolidate two institutional alarm data into a common set of alarm codes. It works well given that alarms from both institutions are from bedside patient monitors of the same vendor and differences in alarm messages are not prominent across different models from this particular vendor. More complex mapping schemes, such as various word embedding techniques[53]-[55], are worth further exploring to push the envelope of SuperAlarm towards generalization even across alarm data of different vendors.

Acknowledgments

This work was supported by the U.S. National Heart, Lung, and Blood Institute (NHLBI) under Grant R01HL128679.

Biographies

RAN XIAO (M'15) received his master's and PhD degree from the School of Electrical and Computer Engineering in the University of Oklahoma in 2010 and 2015. He then finished his postdoctoral training at University of Oklahoma in 2016, University of Southern California in 2017 and University of California San Francisco in 2019 respectively. He is currently an Assistant Professor at Duke University School of Nursing, where he conducts research in the field of biomedical informatics and application of signal processing and machine learning in biomedical research.

Dr. Xiao was a recipient of the Brain-Computer Interface Scholarship from the BCI Society in 2013, and the winner of Jos Willems Early Career Investigator Competition from the International Society for Computerized Electrocardiology in 2018.

DUC DO MD MS received a BS in biological sciences at the University of California, Riverside in 2007 and an MD from the University of California, Los Angeles in 2011. He completed residency in Internal Medicine in 2014 at Stanford University, fellowships in Cardiovascular Diseases and Clinical Cardiac Electrophysiology at the University of California, Los Angeles in 2018 and 2020, respectively. He also completed a Masters in Clinical Research during his cardiology fellowship. He is currently a Clinical Instructor in Cardiac Electrophysiology at the UCLA Cardiac Arrhythmia Center at the David Geffen School of Medicine at the University of California, Los Angeles. His research focuses on utilizing continuous electrocardiographic data to predict cardiac arrests and gain insights into their pathophysiology.

CHENG DING received the B.S. degree in Computer Science from Anhui University in Hefei, China in 2013 and the M.S. degree in software engineering from University of Science and Technology of China, Hefei, China, in 2017, and second M.S. degree in Statistics from University of Arizona, Tucson, USA. He is currently pursuing the Ph.D. degree in Electronical and Computer engineering at Duke University, Durham, USA.

From 2018 to 2020, he was a Research Assistant in Hulab at University of California, San Francisco. His research interest includes electronical health record data management and developing machine learning and deep learning algorithms to detect early stage of clinical endpoint.

KARL MEISEL is an Assistant Professor of Neurology at UCSF and provided direct supervision of residents/fellows admitting patients with neurovascular disease. Also, I have developed an automated online screening survey for disability and depression in stroke clinic patients. As the attending physician on the neurovascular service I teach residents, medical students and fellows. I have collaborated successfully to produce a number of publications including the NIH funded POINT trial. I was the local PI for the RESPECT-ESUS and STROKE-AF trials. My position is the director of the stroke clinic at UCSF in which I evaluate stroke patients who have embolic stroke of undetermined source (ESUS) enrolling qualifying patients into the NIH funded ARCADIA trial as the local PI. Also I am the local PI for PFO closure studies by Abbott and GORE that evaluates the outcome of patients with ESUS treated with PFO closure. My passion is to prevent future strokes by early detection and intervention of risk factors. This passion has led to my effort to develop a PPG monitoring system to detect atrial fibrillation.

RANDALL J. LEE, MD, PhD, Professor, Medicine at the University of California, San Francisco, is a Cardiologist and a Cardiac Electrophysiologist who specializes in the treatment of arrhythmias and prevention of stroke.

He obtained has MD and PhD degree in Pharmacology at the University of California, Los Angeles in 1984, completed his medicine residency at Harbor-UCLA in 1988, concluded his cardiology fellowship at Stanford University in 1991 and finished his clinical training in cardiac electrophysiology at the University of California, San Francisco in 1992.

In addition to the development of devices and techniques for the treatment of arrhythmias and embolic stroke prevention, Dr. Lee has an active cardiac tissue engineering laboratory for myocardial repair/reconstruction.

XIAO HU is Ann Henshaw Gardiner Distinguished Professor of Nursing at Duke University. He also has secondary appointment in departments of Neurology, Surgery, and Biostatistics & Bioinformatics in the School of Medicine and department of Electrical &

Computer Engineering in the Pratt School of Engineering at Duke University. Dr. Hu obtained his bachelor and master's degrees in Biomedical Engineering from University of Electronic Science and Technology in 1996 and 1999, respectively. He started his PhD program in Biomedical Engineering at University of California, Los Angeles in 1999 and completed it in 2004. He became an independent researcher in 2006 at department of Neurosurgery at University of California Los Angeles. Dr. Hu's expertise includes biomedical signal processing, mathematical modeling of cerebral hemodynamics, machine learning, database / informatics systems, and software development. He has been the principal investigator of numerous NIH-funded projects. Dr. Hu has published more than 130 journal papers and has been awarded 9 US patents. He is the editor-in-chief for Physiological Measurement and a standing panel member of NIH/CSR Biomedical Computing and Health Informatics study section.

References

[1]. Meredith C and Edworthy J, "Are there too many alarms in the intensive care unit? An overview of the problems," J. Adv. Nurs, 1995, doi: 10.1046/j.1365-2648.1995.21010015.x.

[2]. Lawless ST, "Crying wolf: False alarms in a pediatric intensive care unit," Crit. Care Med, 1994, doi: 10.1097/00003246-199406000-00017.

[3]. Clifford GD et al., "False alarm reduction in critical care," Physiological Measurement. 2016, doi: 10.1088/0967-3334/37/8/E5.

[4]. Siebig S, Kuhls S, Imhoff M, Gather U, Schölmerich J, and Wrede CE, "Intensive care unit alarms-How many do we need?," Crit. Care Med, 2010, doi: 10.1097/CCM.0b013e3181cb0888.

[5]. Chambrin MC, "Alarms in the intensive care unit: How can the number of false alarms be reduced?," Critical Care. 2001, doi: 10.1186/cc1021.

[6]. Imhoff M and Kuhls S, "Alarm algorithms in critical monitoring," Anesth. Analg, 2006, doi: 10.1213/01.ane.0000204385.01983.61.

[7]. Drew BJ et al., "Insights into the problem of alarm fatigue with physiologic monitor devices: A comprehensive observational study of consecutive intensive care unit patients," PLoS One, vol. 9, no. 10, p. e110274, 2014, doi: 10.1371/journal.pone.0110274. [PubMed: 25338067]

[8]. Graham KC and Cvach M, "Monitor alarm fatigue: Standardizing use of physiological monitors anddecreasing nuisance alarms," Am. J. Crit. Care, 2010, doi: 10.4037/ajcc2010651.

[9]. Sendelbach S, Wahl S, Anthony A, and Shotts P, "Stop the noise: A quality improvement project to decrease electrocardiographic nuisance alarms," Crit. Care Nurse, 2015, doi: 10.4037/ccn2015858.

[10]. Welch J, "An evidence-based approach to reduce nuisance alarms and alarm fatigue," Biomedical Instrumentation and Technology. 2011, doi: 10.2345/0899-8205-45.s1.46.

[11]. Tanner T, "The problem of alarm fatigue," Nurs. Womens. Health, 2013, doi: 10.1111/1751-486X.12025.

[12]. Casey S, Avalos G, and Dowling M, "Critical care nurses' knowledge of alarm fatigue and practices towards alarms: A multicentre study," Intensive Crit. Care Nurs, 2018, doi: 10.1016/j.iccn.2018.05.004.

[13]. Cvach M, "Monitor alarm fatigue: An integrative review," Biomedical Instrumentation and Technology. 2012, doi: 10.2345/0899-8205-46.4.268.

[14]. Sendelbach S, "Alarm Fatigue," Nursing Clinics of North America. 2012, doi: 10.1016/j.cnur.2012.05.009.

[15]. Johnson KR, Hagadorn JI, and Sink DW, "Alarm Safety and Alarm Fatigue," Clinics in Perinatology. 2017, doi: 10.1016/j.clp.2017.05.005.

[16]. Sendelbach S and Funk M, "Alarm fatigue: A patient safety concern," AACN Adv. Crit. Care, 2013, doi: 10.1097/NCI.0b013e3182a903f9.

[17]. Ruskin KJ and Hueske-Kraus D, “Alarm fatigue: Impacts on patient safety,” Current Opinion in Anaesthesiology. 2015, doi: 10.1097/ACO.0000000000000260.

[18]. Horkan AM, “Alarm fatigue and patient safety.,” Nephrol. Nurs. J, 2014.

[19]. Eerikainen LM, Vanschoren J, Rooijakkers MJ, Vullings R, and Aarts RM, “Reduction of false arrhythmia alarms using signal selection and machine learning,” Physiol. Meas, 2016, doi: 10.1088/0967-3334/37/8/1204.

[20]. Behar J, Oster J, Li Q, and Clifford GD, “ECG signal quality during arrhythmia and its application to false alarm reduction,” IEEE Trans. Biomed. Eng, 2013, doi: 10.1109/TBME.2013.2240452.

[21]. Clifford GD et al., “Using the blood pressure waveform to reduce critical false ECG alarms,” in Computers in Cardiology, 2006.

[22]. Li Q and Clifford GD, “Signal quality and data fusion for false alarm reduction in the intensive care unit,” J. Electrocardiol, 2012, doi: 10.1016/j.jelectrocard.2012.07.015.

[23]. Masino AJ et al., “Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data,” PLoS One, 2019, doi: 10.1371/journal.pone.0212665.

[24]. O’Brien C et al., “Development, Implementation, and Evaluation of an In-Hospital Optimized Early Warning Score for Patient Deterioration,” MDM Policy Pract, 2020, doi: 10.1177/2381468319899663.

[25]. Parshuram CS, Hutchison J, and Middaugh K, “Development and initial validation of the Bedside Paediatric Early Warning System score,” Crit. Care, 2009, doi: 10.1186/cc7998.

[26]. Hill K, “National Early Warning Score,” Nurs. Crit. Care, 2012, doi: 10.1111/j.1478-5153.2012.00540_3.x.

[27]. Salas-Boni R, Bai Y, and Hu X, “Cumulative Time Series Representation for Code Blue prediction in the Intensive Care Unit.,” AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci, vol. 2015, pp. 162–167, 2015.

[28]. Bai Y et al., “Is the Sequence of SuperAlarm Triggers More Predictive Than Sequence of the Currently Utilized Patient Monitor Alarms?,” IEEE Trans. Biomed. Eng, vol. 64, no. 5, pp. 1023–1032, 2017, doi: 10.1109/TBME.2016.2586443. [PubMed: 27390164]

[29]. Hu X et al., “Predictive combinations of monitor alarms preceding in-hospital code blue events,” J. Biomed. Inform, vol. 45, no. 5, pp. 913–921, 2012, doi: 10.1016/j.jbi.2012.03.001. [PubMed: 22465785]

[30]. Bai Y et al., “Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction,” J. Biomed. Inform, vol. 53, no. November, pp. 81–92, 2015, doi: 10.1016/j.jbi.2014.09.006. [PubMed: 25240252]

[31]. Agrawal R, Imieliński T, and Swami A, “Mining Association Rules Between Sets of Items in Large Databases,” ACM SIGMOD Rec, 1993, doi: 10.1145/170036.170072.

[32]. Tsai CJ, Lee CI, and Yang WP, “A discretization algorithm based on Class-Attribute Contingency Coefficient,” Inf. Sci. (Ny), 2008, doi: 10.1016/j.ins.2007.09.004.

[33]. Lewis PAW and Shedler GS, “SIMULATION OF NONHOMOGENEOUS POISSON PROCESSES BY THINNING.,” Nav. Res. Logist. Q, 1979, doi: 10.1002/nav.3800260304.

[34]. Burdick D, Calimlim M, Flannick J, Gehrke J, and Yiu T, “MAFIA: A maximal frequent itemset algorithm,” IEEE Trans. Knowl. Data Eng, 2005, doi: 10.1109/TKDE.2005.183.

[35]. Burdick D, Calimlim M, and Gehrke J, “MAFIA: A maximal frequent itemset algorithm for transactional databases,” Proc. - Int. Conf. Data Eng, 2001, doi: 10.1109/ICDE.2001.914857.

[36]. Akobeng AK, “Understanding diagnostic tests 1: Sensitivity, specificity and predictive values,” Acta Paediatrica, International Journal of Paediatrics. 2007, doi: 10.1111/j.1651-2227.2006.00180.x.

[37]. AA CZ Mooney, and Duval RD, “Bootstrapping: A Nonparametric Approach to Statistical Inference.,” J. Am. Stat. Assoc, 1994, doi: 10.2307/2290969.

[38]. Rice TK, Schork NJ, and Rao DC, “Methods for Handling Multiple Testing,” Advances in Genetics. 2008, doi: 10.1016/S0065-2660(07)00412–9.

[39]. A1sgari S, Xu P, Bergsneider M, and Hu X, “A subspace decomposition approach toward recognizing valid pulsatile signals,” Physiol. Meas, 2009, doi: 10.1088/0967-3334/30/11/006.

[40]. Aboukhalil A, Nielsen L, Saeed M, Mark RG, and Clifford GD, “Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform,” J. Biomed. Inform, 2008, doi: 10.1016/j.jbi.2008.03.003.

[41]. Zong W, Moody GB, and Mark RG, “Reduction of false arterial blood pressure alarms using signal quality assessment and relationships between the electrocardiogram and arterial blood pressure,” Med. Biol. Eng. Comput, 2004, doi: 10.1007/BF02347553.

[42]. Asgari S, Bergsneider M, and Hu X, “A robust approach toward recognizing valid arterial-blood-pressure pulses,” in IEEE Transactions on Information Technology in Biomedicine, 2010, doi: 10.1109/TITB.2009.2034845.

[43]. Downing NL et al., “Electronic health record-based clinical decision support alert for severe sepsis: A randomised evaluation,” BMJ Qual. Saf, 2019, doi: 10.1136/bmjqs-2018-008765.

[44]. Subbe CP, “Validation of a modified Early Warning Score in medical admissions,” QJM, 2001, doi: 10.1093/qjmed/94.10.521.

[45]. Finlay GD, Rothman MJ, and Smith RA, “Measuring the modified early warning score and the Rothman Index: Advantages of utilizing the electronic medical record in an early warning system,” J. Hosp. Med, 2014, doi: 10.1002/jhm.2132.

[46]. Bedoya AD, Clement ME, Phelan M, Steorts RC, O’Brien C, and Goldstein BA, “Minimal Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration,” Crit. Care Med, 2019, doi: 10.1097/CCM.0000000000003439.

[47]. Gardner-Thorpe J, Love N, Wrightson J, Walsh S, and Keeling N, “The value of Modified Early Warning Score (MEWS) in surgical inpatients: A prospective observational study,” Ann. R. Coll. Surg. Engl, 2006, doi: 10.1308/003588406×130615.

[48]. Hu X, “An algorithm strategy for precise patient monitoring in a connected healthcare enterprise,” npj Digit. Med, 2019, doi: 10.1038/s41746-019-0107-z.

[49]. Hu X, Do D, Bai Y, and Boyle NG, “A case-control study of non-monitored ECG metrics preceding in-hospital bradyasystolic cardiac arrest: Implication for predictive monitor alarms,” J. Electrocardiol, vol. 46, no. 6, pp. 608–615, 2013, doi: 10.1016/j.jelectrocard.2013.08.010. [PubMed: 24034301]

[50]. Hock Ong ME et al., “Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score,” Crit. Care, 2012, doi: 10.1186/cc11396.

[51]. Kim J, Chae M, Chang H-J, Kim Y-A, and Park E, “Predicting Cardiac Arrest and Respiratory Failure Using Feasible Artificial Intelligence with Simple Trajectories of Patient Data,” J. Clin. Med, 2019, doi: 10.3390/jcm8091336.

[52]. Kwon JM, Lee Y, Lee Y, Lee S, and Park J, “An algorithm based on deep learning for predicting in-hospital cardiac arrest,” J. Am. Heart Assoc, 2018, doi: 10.1161/JAHA.118.008678.

[53]. Si Y, Wang J, Xu H, and Roberts K, “Enhancing clinical concept extraction with contextual embeddings,” J. Am. Med. Informatics Assoc, 2019, doi: 10.1093/jamia/ocz096.

[54]. Li Y and Yang T, “Word Embedding for Understanding Natural Language: A Survey,” 2018.

[55]. Ling Y, “Methods and Techniques for Clinical Text Modeling and Analytics,” ProQuest Diss. Theses, 2017.

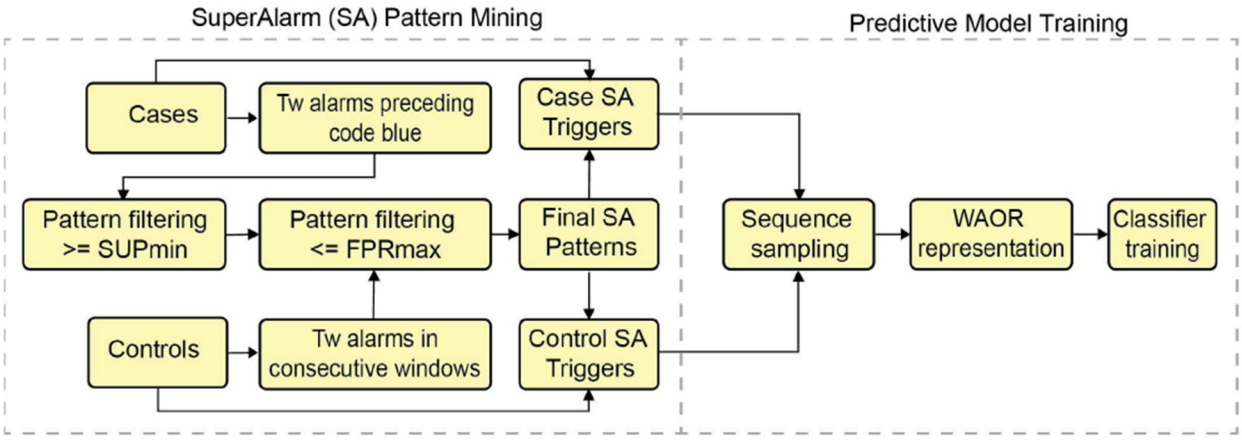

Fig. 1.

The flowchart for SuperAlarm model training. The left panel shows procedures for mining SuperAlarm patterns. The right panel shows procedures for training the predictive model using trigger sequences of SuperAlarm patterns.

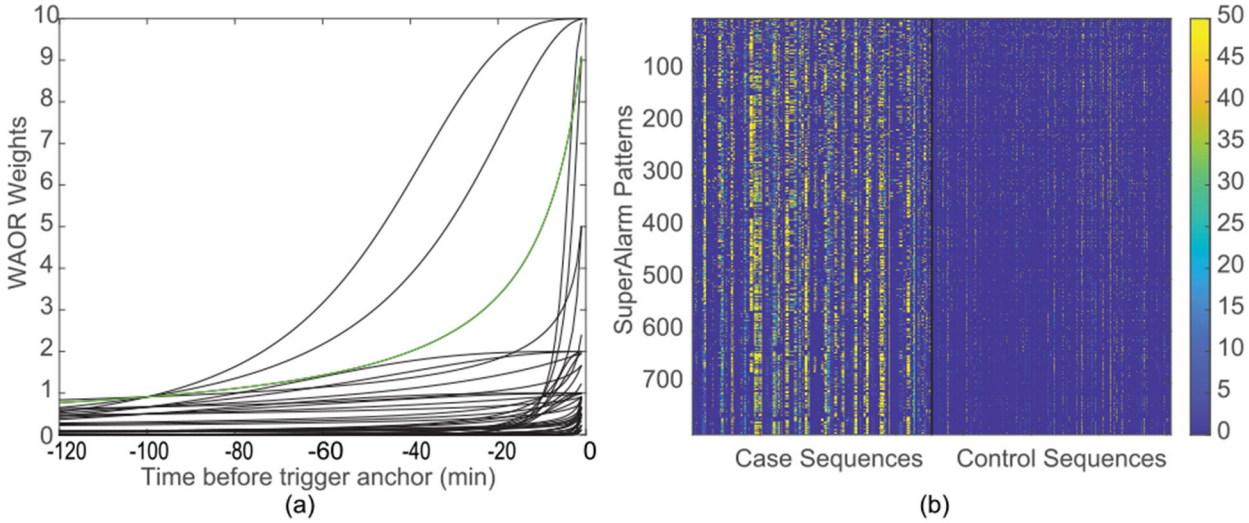

Fig. 2.

Profiles and feature visualization of WAOR sequence representation algorithm. (a) shows visualization of WAOR profiles evaluated for predictive model training. The green curve designates the final WAOR profile determined during the training process; (b) shows WAOR weights of case and control sequences separated by the black vertical line. Rows denote different SuperAlarm patterns and columns representing different trigger sequences.

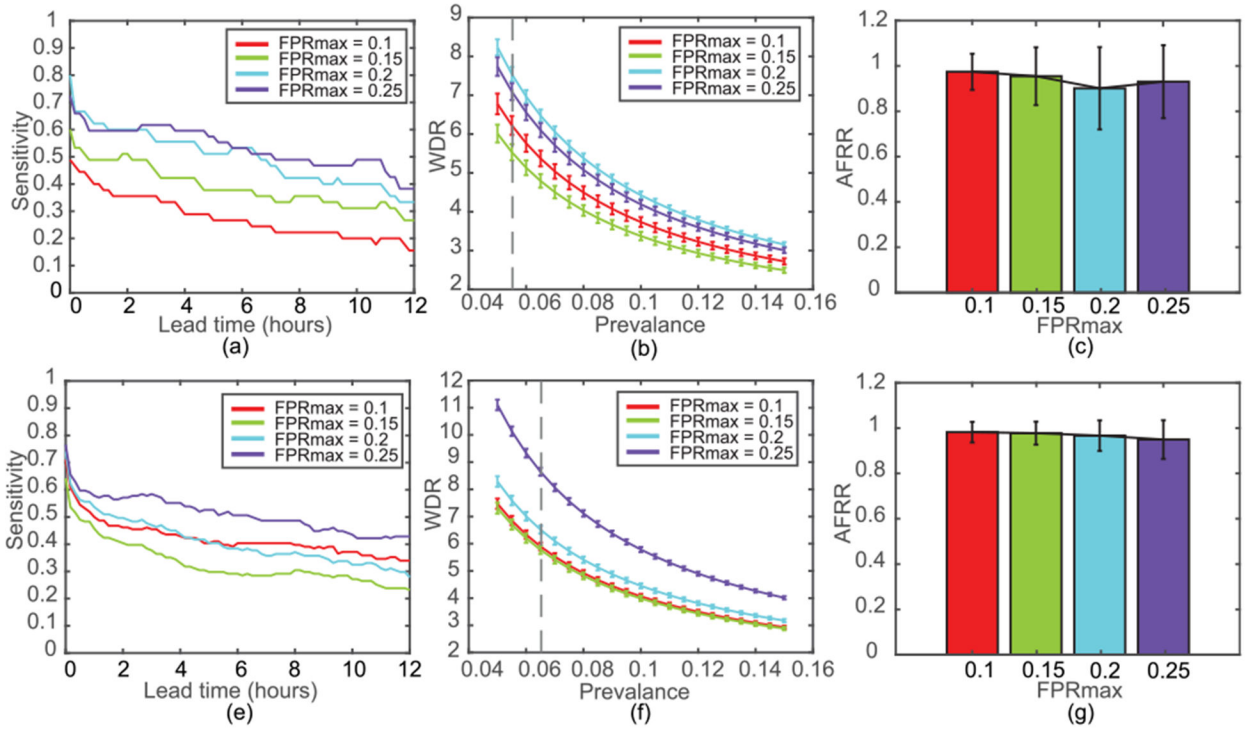

Fig. 3.

SuperAlarm performance on internal and external test sets. (a)-(c) show sensitivity, WDR and AFRR, respectively, from models with various FPRmax tested on an internal test set; (e)-(f) show sensitivity, WDR and AFRR, respectively, from models with various FPRmax tested on an external test set.

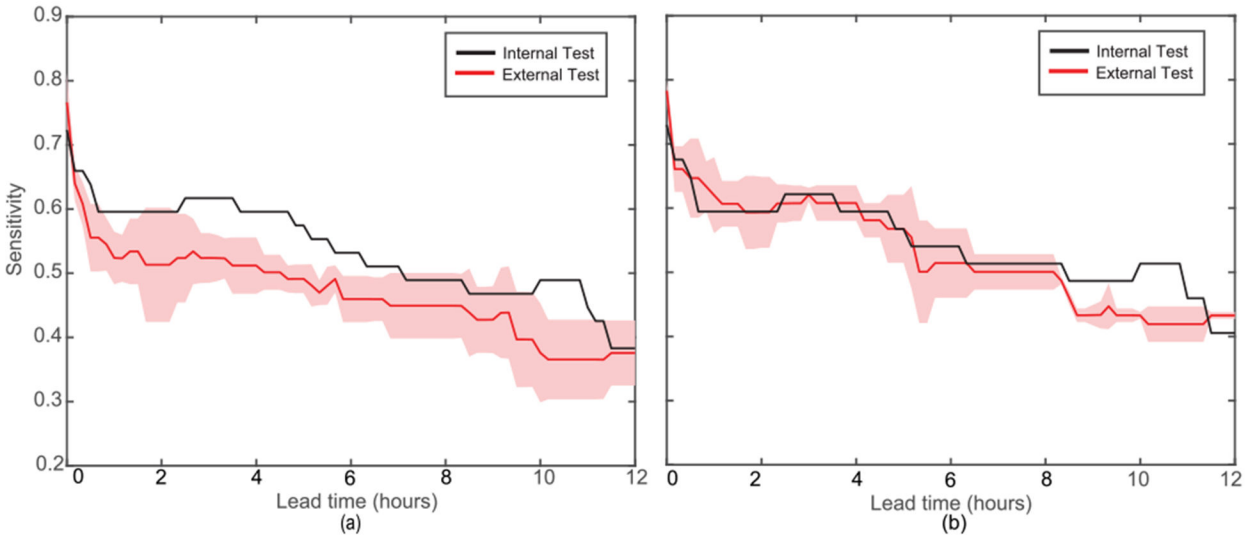

Fig. 4.

Statistical comparison of sensitivity along lead time between internal and bootstrapping-based external tests. (a) shows comparison of cross-institutional sensitivity in predicting all in-hospital code blue events; (b) shows a comparison of cross-institutional sensitivity in predicting cardiopulmonary arrest, the dominant subtype of code blue.

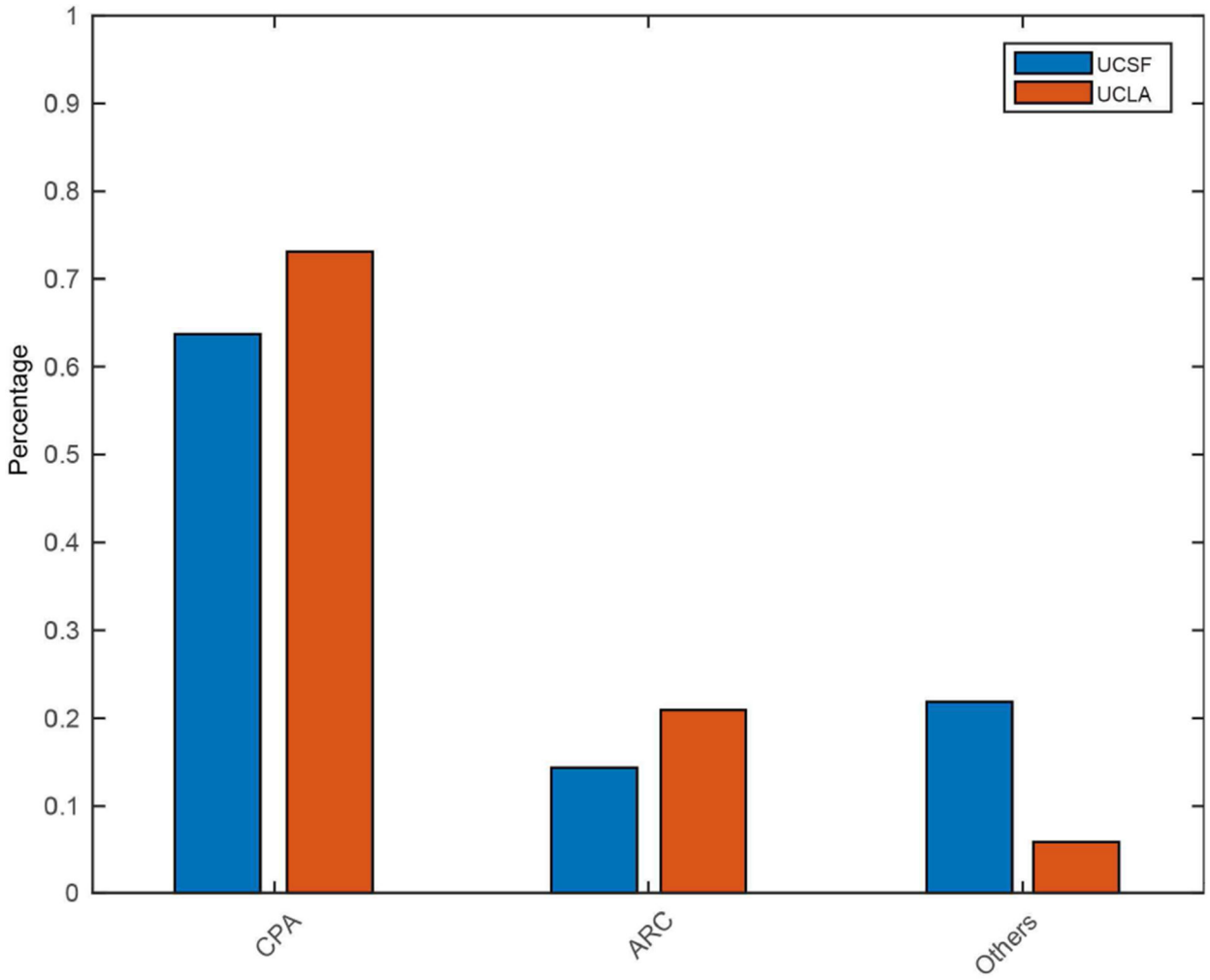

Fig. 5.

Distribution of code blue subtypes in two institutions. CPA, cardiopulmonary arrest; ARC, acute respiratory compromise; Others, other medical emergencies.

TABLE I.

Hyperparameters selected during SuperAlarm pattern mining

FPRmax	Tw	SUPmin	# SAPattern
0.1	1 hour	0.1	377
0.15	2 hours	0.1	619
0.2	2 hours	0.1	691
0.25	1.5 hours	0.1	798

TABLE II.

Most frequent SuperAlarm patterns at each cardinality.

Cardinality	SA Patterns based on FPRmax of 0.25	12h TPR
2	“BP SYS HI: [179.5, 256.5]”, “BP DIA LO: [35.5,45.5]”	0.231
3	“BP RATE LO: [,33.5]”, “BP SYS LO: [70.5, 82.5]”, “PVC”	0.280
4	“BP DIA LO: [35.5,45.5]”, “COUPLET”, “RESP HI: [,113.5]”, “PVC”	0.253
5	“BP DIA LO: [35.5,45.5]”, “BP DIA LO: [45.5, 69.5]”, “BP SYS LO: [82.5, 199.5]”,” BP MEAN LO: [55.5, 116.5]”, “RESPHI: [,113.5]”	0.253
6	“BP SYS LO: [70.5, 82.5]”, “BP SYS LO: [82.5, 199.5]”, “BP MEAN LO: [55.5,116.5]”, “BP DIA LO: [35.5, 45.5]”, “RESP HI: [,113.5]”, “PVC”	0.220
7	“BP MEAN LO: [19.5, 51.5]”, “BP MEAN LO: [51.5, 55.5]”, “BP MEAN LO: [55.5, 116.5]”, “BP DIA LO: [35.5, 45.5]”, “BP SYS LO: [26.5, 70.5]”, “BP SYS LO: [70.5, 82.5]”, “PVC”	0.209

TABLE III.

Offline performance of SuperAlarm on internal test set.

FPRmax	AUC (%)	Accu. (%)	Sen. (%)	Spec. (%)	PPV (%)	NPV (%)	F1 (%)
0.1	79.16	75.59	67.40	78.41	51.91	87.44	58.65
0.15	82.01	77.46	65.73	82.15	59.55	85.71	62.49
0.2	83.56	70.39	86.38	64.44	47.49	92.71	61.28
0.25	85.96	79.40	81.13	78.89	53.46	93.33	64.45