Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction

Yong Bai, Duc H. Do, Patricia Rae Eileen Harris, Daniel Schindler, Noel G. Boyle, Barbara J. Drew, Xiao Hu

Department of Bioengineering, University of California, Los Angeles, CA, United States
UCLA Cardiac Arrhythmia Center, David Geffen School of Medicine, University of California, Los Angeles, CA, United States
Department of Physiological Nursing, University of California, San Francisco, CA, United States
Department of Neurosurgery, University of California, San Francisco, CA, United States
Institute for Computational Health Sciences, University of California, San Francisco, CA, United States
UCB/UCSF Graduate Group in Bioengineering, University of California, San Francisco, CA, United States

ABSTRACT

Patient monitors in modern hospitals have become ubiquitous but they generate an excessive number of false alarms causing alarm fatigue. Our previous work showed that combinations of frequently co-occurring monitor alarms, called SuperAlarm patterns, were capable of predicting in-hospital code blue events at a lower alarm frequency. In the present study, we extend the conceptual domain of a SuperAlarm to incorporate laboratory test results along with monitor alarms so as to build an integrated data set to mine SuperAlarm patterns. We propose two approaches to integrate monitor alarms with laboratory test results and use a maximal frequent itemsets mining algorithm to find SuperAlarm patterns. Under an acceptable false positive rate $FPR_{max}$, optimal parameters including the minimum support threshold and the length of time window for the algorithm to find the combinations of monitor alarms and laboratory test results are determined based on a 10-fold cross-validation set. SuperAlarm candidates are generated under these optimal parameters. The final SuperAlarm patterns are obtained by further removing the candidates with false positive rate $> FPR_{max}$. The performance of SuperAlarm patterns are assessed using an independent test data set. First, we calculate the sensitivity with respect to prediction window and the sensitivity with respect to lead time. Second, we calculate the false SuperAlarm ratio (ratio of the hourly number of SuperAlarm triggers for control patients to that of the monitor alarms, or that of regular monitor alarms plus laboratory test results if the SuperAlarm patterns contain laboratory test results) and the work-up to detection ratio, WDR (ratio of the number of patients triggering any SuperAlarm patterns to that of code blue patients triggering any SuperAlarm patterns). The experiment results demonstrate that when varying $FPR_{max}$ between 0.02 and 0.15, the SuperAlarm patterns composed of monitor alarms along with the last two laboratory test results are triggered at least once for [56.7–93.3%] of code blue patients within an 1-h prediction window before code blue events and for [43.3–90.0%] of code blue patients at least 1-h ahead of code blue events. However, the hourly number of these SuperAlarm patterns occurring in control patients is only [2.0–14.8%] of that of regular monitor alarms with WDR varying between 2.1 and 6.5 in a 12-h window. For a given $FPR_{max}$ threshold, the SuperAlarm set generated from the integrated data set has higher sensitivity and lower WDR than the SuperAlarm set generated from the regular monitor alarm data set. In addition, the McNemar’s test also shows that the performance of the SuperAlarm set from the integrated data set is significantly different from that of the SuperAlarm set from the regular monitor alarm data set. We therefore conclude that the SuperAlarm patterns generated from the integrated data set are better at predicting code blue events.

© 2014 Elsevier Inc. All rights reserved.

Please cite this article in press as: Bai Y et al. Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.09.006
1. Introduction

With technologic advances in medical devices over the past few decades, life-saving patient monitoring systems have become ubiquitous in modern hospitals [1]. Alarms annunciated by the monitoring systems are expected to alert caregivers to either changes in monitored physiological parameters of a patient or device malfunction, and to enhance quality of care and patient safety by detection of any abnormality [2].

In traditional monitor algorithms, an alarm is triggered immediately when the value of the monitored parameter exceeds or falls below the preset threshold [3]. Due to the lack of a standard for default threshold setting [4], this threshold-based algorithm is intentionally set to have high sensitivity in order to capture the greatest percentage of clinically significant events [5,6]. As a consequence, there is low specificity and numerous alarms occur (about 700 alarms per patient per day [7]) and up to 99% of them are false alarms and nuisance (or false positive) alarms with no clinical relevance [2,5,7–10]. Excessive false and nuisance alarms may compromise the quality of patient care and cause unexpected alarm-related deaths in hospitals [12]. The alarm hazard has been ranked as the “TOP 1” technology hazard for 2014 by the Emergency Care Research Institute (ECRI) [13].

Many studies have focused on addressing the alarm fatigue problem. Descriptions of many such algorithms were provided in reviews [1,14]. For instance, Zong et al. [15] proposed an algorithm for reducing false arterial blood pressure (ABP) alarms by evaluating signal quality of ABP and the relationship between electrocardiogram (ECC) and ABP using fuzzy logic approach. Similarly, Aboukhail et al. [16] reduced false critical ECC arrhythmia alarms using morphological and timing information derived from the ABP waveforms. Lastly, Li et al. [17] used a machine learning technique and data fusion method to reduce false arrhythmia alarms by combining signal quality and physiological metrics derived from the waveforms of ECC, photoplethysmograph, and optionally, ABP. We applied pattern recognition methods to reduce false intracranial pressure (ICP) alarms using the morphological waveform features extracted from the ICP signal [18,19]. These approaches were developed to manage individual alarm types and further validation is needed to ensure that no true alarm is suppressed before their implementations by monitor vendors. Additionally, true alarms not suppressed by these approaches were designed to detect abnormalities after they occur, not to detect patient deterioration. Therefore, they are at best able to support a reactive patient care practice rather than a predictive one.

To detect patient deterioration, especially outside intensive care units, several score-based systems have been developed based on multiple parameters. The modified early warning score (MEWS) [20], for instance, was a simple tool to produce a fusion score based on the summation of an individual score assigned to each of five physiological parameters: systolic blood pressure (SysBP), respiratory rate (RR), pulse rate, temperature and patient consciousness. For each parameter, the greater the degree of deviation from the normal range, the larger the individual score assigned. However, the schema for score assignment was designed empirically [21]. Biosign [22,23] was another algorithm to generate a patient status index (PSI) by fusing five vital signs: heart rate (HR), respiratory rate (RR), blood pressure (BP), temperature and arterial oxygen saturation (SpO2). It used a multivariate Gaussian probabilistic model for the distribution of these vital signs for patients without crisis events. A patient crisis event was detected when these vital signs had a small probability according to this distribution estimated from a training data set. Rothman et al. [24] developed a system to calculate a patient acuity metric, called the Rothman Index (RI), to evaluate the risk of patient deterioration using vital signs, laboratory test results, indicators of cardiac rhythms, and nursing assessments. This approach was based on empirical accumulation of relative risks of its component variables in determining patient mortality after 1 year discharge from the hospital. Machine learning-based methods have also been proposed to detect patient deterioration. For instance, Clifton et al. [25] compared Gaussian mixture model (GMM) and support vector machine (SVM) with HR, RR, SpO2, and SysBP as input. Tarassenko et al. [26] developed a centile-based early warning score system based on statistical properties of the vital signs (HR, RR, SpO2 and SysBP) to identify deteriorating patients. Scores were determined when the statistical value of vital sign fell into a certain range of centile.

It can be argued that those algorithms presented above for detection of patient deterioration introduce additional alarms or alerts without providing direct relief of the existing alarm fatigue problem. A potentially more desirable approach would incorporate patient monitor alarms and physiological signals from patient monitors. The idea to include monitor alarms as predictors of patient deterioration detection models has been tested by our group. In our previous paper [27], we proposed a novel data-driven approach using raw streaming alarm data to: (1) identify patterns that were combined with different monitor alarms using in-hospital code blue events; (2) select those patterns that occurred sufficiently often preceding code blue events but rarely in control patients; (3) empirically define and determine the optimal length of time window for the selected patterns; (4) assess the temporal characteristics of these patterns such as the sensitivity with respect to prediction window; and then (5) based on these factors, evaluate the performance of these patterns, which we called SuperAlarm patterns, under varying acceptable false positive rates. Because a SuperAlarm trigger necessarily requires simultaneous triggering of different alarms, it therefore has the potential to reduce alarm frequency.

In the present study, we follow the general framework we have previously proposed [27] and describe how we extend the conceptual domain of a SuperAlarm to incorporate laboratory test results as an additional source to compose SuperAlarm patterns. To do so, we propose several new methods so as to tackle complicating factors that arise when one incorporate non-streaming data (e.g., patients with very sparse data). We also address the need to exclude “crisis” alarms that clinicians would consider to be “non-brainers” such as asystole. Specifically, we first explore a Non-Homogenous Poisson Process (NHPP) to model the occurrence rate of monitor alarms and obtain an objective threshold to exclude code blue patients with unexpectedly small number of monitor alarms preceding code blue events. We then develop two approaches to integrate laboratory test results with monitor alarms. We apply a new algorithm to discover SuperAlarm candidate patterns occurring frequently before code blue events. These candidate patterns are composed of combinations of maximal number of monitor alarms and laboratory test results with occurrence rate greater than a support threshold. The candidate patterns are further filtered out if their false positive rates are greater than an acceptable false positive rate \( FPR_{max} \), resulting in the final SuperAlarm patterns. By construction, these patterns are less redundant compared to those determined by the techniques of mining frequent itemsets (FI) or closed frequent itemsets (CFI) used in our previous work.

2. Methods

Fig. 1 illustrates the flowchart of the proposed algorithm to discover SuperAlarm patterns. Key steps of this process are described in the following sections.
The average number of alarms per patient over a specific period, $\bar{n}$, will be the estimated number of alarms occurring at $t$ over time interval $[0, T]$ such that

$$\bar{n} = \int_0^T \mu(t) dt = \int_0^T e^{x+Y} dt$$

(1)

The time interval $[0, T]$ is divided into $N$ subintervals $[\frac{k-1}{N}, \frac{k}{N}]$, $1 \leq k \leq N$. Let $\lambda_k$ be the average number of alarms per patient over the subinterval $k$. We then use the generalized linear model (GLM) to estimate the parameters $\lambda$ and $Y$.

The estimated number of alarms over $T_w$ is given by

$$\hat{n} = \int_0^{T_w} \mu(t) dt = \int_0^{T_w} e^{x+Y} dt$$

(2)

95% interval of $\hat{n}$ is $[n_{\text{lower}}, n_{\text{upper}}]$. Thus, the minimum-alarm-count-threshold over the $T_w$ is defined as

$$N_{\text{minCount}} = [n_{\text{lower}}]$$

(3)

where $x$ is the maximum integral number that is not greater than $x$.

We exclude those code blue patients whose number of alarms within a $T_w$-long time window preceding code blue events is less than the $N_{\text{minCount}}$ threshold. Regular monitor alarms from the rest of patients constitute the $\text{Alarm}$ data set.

### 2.3. Integration of monitor alarms with laboratory test results

Two approaches are proposed to integrate monitor alarms with laboratory test results. Using the first approach as illustrated in panel A of Fig. 2, we integrate the latest abnormal result of each type of laboratory test with the array of monitor alarms within a $T_w$-long window. We select abnormal laboratory test results from our data set based on the associated flags reported by the electronic medical record (EMR) system. There are five flags for laboratory test results: $HH$ (extremely high), $H$ (high), $L$ (low), $LL$ (extremely low), and $N$ (normal). The abnormality flags for a given laboratory test result include $HH, H, L, LL$, and $N$. In this way, we ignore the numeric value of an abnormal laboratory test result and adopt the following representation: "[Test Name] [Abnormality]." For instance, if the laboratory test result "WBC" was flagged by $H$, then it would be represented as "WBC H." It can be seen from Fig. 2(A) that $LA$ and $LB$ represent arrays of abnormal results from two different laboratory tests for a given patient. We will select $LA_1$ and $LA_2$ and integrate them with monitor alarms as they are the latest results of the laboratory tests to fall outside the time window specified by $T_w$.

In the second approach, we use the difference between the last two results of a laboratory test before a $T_w$-long window as a laboratory test trigger to be integrated with monitor alarms (panel B of Fig. 2). As each laboratory test result can be indicated by one of the five flags $HH, H, L, LL$, and $N$, there will be 25 possible triggers for a given laboratory test, which we called delta laboratory test results: $HH \rightarrow HH$, $HH \rightarrow H$, $HH \rightarrow L$, $HH \rightarrow LL$, $HH \rightarrow N$, ..., $N \rightarrow HH$, $N \rightarrow H$, $N \rightarrow L$, $N \rightarrow LL$, $N \rightarrow I$. For instance, if the last two results of laboratory test "Hemoglobin" were flagged by $N$ and $L$, then the delta laboratory test result would be represented as "Hemoglobin $N \rightarrow L$". From Fig. 2(B), we can see that $LA$ represents an array of results from a laboratory test for a given patient. $LA_1$ and $LA_2$ will be selected and integrated with monitor alarms within a $T_w$-long window since $LA_1$ and $LA_2$ are the two latest results for laboratory test $LA$ with respect to $T_w$.

Based on these two approaches, we create two extended data sets: the $\text{Ab Lab + Alarm}$ data set, which is composed by the $\text{Alarm}$ data set integrated with the abnormal laboratory test results, and the $\text{Delta Lab + Alarm}$ data set, which consists of the $\text{Alarm}$ data set integrated with the delta laboratory test results.
2.4. Discovery of SuperAlarm patterns

To facilitate discovery of SuperAlarm patterns, we first encode parametric monitor alarms by discretizing their numeric values using the Class-Attribute Contingency Coefficient (CACC) algorithm [28]. The CACC algorithm is a supervised discretization algorithm to generate intervals for given numeric attributes by finding the cutting points. It takes the contingency coefficient into account to measure the strength of dependence between individual attribute and classes. Therefore, the CACC algorithm allows us to utilize data from code blue patients and control patients to generate high-quality discretization schemes for parametric monitor alarms with the best correlation between these alarms and the type of patients (i.e., code blue patients and control patients). Laboratory test results do not need to be encoded since they are not represented with numeric values. The integrated data set of laboratory test results with encoded monitor alarms within \( T_w \)-long window preceding code blue events is then used to mine maximal frequent itemsets (MFI), i.e., SuperAlarm candidates.

Definition 1. Support of an itemset: The support of an itemset is defined as the proportion of code blue patients in the data set who contain the itemset.

Definition 2. Frequent Itemsets (FI) [29–32]: An itemset is frequent if its support is not less than a user-specified threshold of minimum support (i.e., \( \min_{sup} \)).

Definition 3. Maximal Frequent Itemsets (MFI) [33–36]: An itemset is maximally frequent if none of its supersets is a frequent itemset. A superset of an itemset is an extension of the itemset.

It should be noted that the following relationship holds between MFI and FI: \( MFI \subseteq FI \). Classic Apriori-based methods mining FI employ a strategy of breadth-first traversal of the search space to find support information for all \( k \)-itemsets \((k = 1, 2, 3,...)\). This method scans all \( 2^k - 2 \) subsets of each \( k \)-itemset to determine whether or not the itemset is frequent based on the Apriori-principle, stating that the superset of any non-FI set is still a non-FI set [30]. Apriori-based method is computationally expensive when the dataset is huge or the frequent itemsets are very long [33,34]. A different method called maximal frequent itemset algorithm (MAFIA) was proposed and it overcame this shortcoming [35].

MAFIA is a new algorithm for maximal frequent itemsets (MFI) mining using depth-first traversal on a lexicographic itemset lattice. Each node on the lattice includes head and tail. The head contains an itemset identifying the node while the tail contains frequent extensions of items lexicographically greater than any items of the head. In the process of depth-first traversal, each item in the node's tail is determined and counted as a 1-extension. According to the Apriori-principle, the traversal process will stop if the support of \( \{\text{nodehead}\} \cup \{1 \text{ - extension}\} \) is less than a user-specified \( \min_{sup} \) threshold. A candidate itemset will be added into MFI set if no superset of this candidate itemset exists in the MFI set. Three pruning strategies are applied to reduce the search space. These include: (1) parent equivalence pruning (PEP); (2) frequent head union tail pruning (FHUT); and (3) head union tail pruning (HUTMF). MAFIA employs vertical bitmaps to represent data and uses an adaptive compression technique to enhance the performance. A vertical bitmap is a column layout to represent the patients for an itemset in the data set, and a bit in a bitmap is used to indicate whether or not the corresponding itemset appears in a given patient. For example, if patient \( i \) has itemset \( j \), then bit \( i \) of the bitmap for itemset \( j \) is set to 1, otherwise, the bit is set to 0. Assume that bitmap \( T \) is a vertical bitmap for itemset \( T \) and bitmap \( S \) for itemset \( S \), then the vertical bitmap for itemset \( T \cup S \), bitmap \( T \cup S \), is defined as bitwise-AND (bitmap \( T \), bitmap \( S \)).

In order to utilize MAFIA to mine MFI, we first build a matrix \( B \) to represent laboratory test results and encoded monitor alarms extracted within \( T_w \)-long window preceding code blue events. \( B = \{x_{ij}\} \) is a \( M \times N \) matrix, where \( M \) is the number of code blue patients and \( N \) is the number of encoded monitor alarms and laboratory test results \((1 \leq i \leq M, 1 \leq j \leq N)\). \( x_{ij} = 0 \) if the \( j \)-th patient does not have the \( j \)-th alarm or laboratory test result, otherwise \( x_{ij} = 1 \). In other words, the \( j \)-th column of \( B \) represents a vertical bitmap for the \( j \)-th alarm or laboratory test result in the data set. The matrix \( B \) is then input into MAFIA under the user-specified \( \min_{sup} \) threshold. As the process of searching goes down the lattice, the head of the node on the lattice grows longer. Due to the sparseness of bitmap especially at the lower support levels, MAFIA compresses...
the bitmap by removing the bit for patient $P$ from itemset $X$ if $P$ does not contain $X$ because MAFIA only needs information about the patients who contain the itemset $X$ to count the support of the subtree rooted at node $n$. MAFIA employs an adaptive compression scheme to determine when to compress the bitmap. In the meanwhile, the three pruning strategies are applied to remove non-maximal sets and therefore reduce the search space. MAFIA adopts the progressive focusing technique to determine whether or not the extracted maximal frequent itemsets are complete. The details of MAFIA can be found in [35]. MAFIA outputs MFI which is a set of patterns consisting of maximal potential components of laboratory test results and monitor alarms.

2.5. Evaluation of SuperAlarm patterns

We evaluate the SuperAlarm patterns by performing both offline and simulated online analysis. Monitor alarms and laboratory test results from a randomly selected 20% of both code blue patients and control patients compose an independent test data set for the simulated online analysis. Those from the remaining 80% of both groups of patients constitute the training data set that is used to build a 10-fold cross-validation set (10-fold CV set) in the offline analysis phase. Optimal parameters of the proposed algorithm are determined based on the performance of the SuperAlarm candidates generated by MAFIA from the 10-fold CV set. The final SuperAlarm set is then generated from the whole training data set under the optimal parameters. This final SuperAlarm set is eventually employed to perform simulated online analysis.

2.5.1. Offline analysis to determine optimal algorithm parameters and generate the final SuperAlarm set

To find the final SuperAlarm patterns, we determine the optimal values of algorithm parameters of $T_w$-long time window and minimum support threshold $\text{min} \_\text{sup}$. This is done by performing cross-validation analysis. According to the integration approaches mentioned in Section 2.3, we extract monitor alarms and laboratory test results within $T_w$-long window preceding code blue events from the first nine folds of the 10-fold CV set. MAFIA is employed to generate SuperAlarm candidates from this extracted data set under a user-specified $\text{min} \_\text{sup}$ threshold. These SuperAlarm candidates are then applied to the first nine folds of the 10-fold CV set for control patients to calculate false positive rate (FPR) values for each of the SuperAlarm candidates. FPR of a SuperAlarm pattern is defined as the percentage of $T_w$-long windows that trigger this pattern in control patients. This is achieved by partitioning the training data set for control patients into consecutive 4-h windows from the beginning of monitoring to the end. A $T_w$-long window is randomly picked within each of these 4-h windows. Laboratory test results and monitor alarms within the $T_w$-long window are used to determine whether a SuperAlarm pattern is triggered, and thereby the FPR of the SuperAlarm pattern is obtained. A SuperAlarm candidate will be removed if it has FPR value greater than a given threshold.

After removing the disqualified SuperAlarm candidates, we apply the rest of SuperAlarm patterns to the remaining one fold of the 10-fold CV set to obtain a pair of values of true positive rate (TPR) and false positive rate (FPR). TPR is defined as the percentage of code blue patients who trigger at least one of SuperAlarm candidates within a $T_w$-long window. FPR here is calculated in terms of percentage of $T_w$-long windows that trigger any of the SuperAlarm patterns in control patients. Varying the threshold will lead to various pairs of TPR and FPR, and hence a receiver operation characteristic (ROC) curve can be generated. This process is repeated for each of the 10 folds, resulting in 10 ROC curves. The final ROC curve is obtained by averaging the 10 ROC curves under a given algorithm parameter combination of $T_w$-long window and $\text{min} \_\text{sup}$.

Given an acceptable false positive rate $\text{FPR}_{\text{max}}$, the optimal values for the parameters of $T_w$ and $\text{min} \_\text{sup}$ are determined by choosing the one with maximal TPR value across all algorithm parameter combinations while possessing FPR value less than $\text{FPR}_{\text{max}}$. Under the optimal algorithm parameter combination, MAFIA is applied again to the whole training data to discover the complete SuperAlarm candidates. The whole training data set is created by coalescing the 10-fold CV data set into one single set. These complete SuperAlarm candidates are further refined to generate final SuperAlarm patterns by filtering out those patterns whose FPR values are greater than $\text{FPR}_{\text{max}}$.

2.5.2. Simulated online analysis

After discovering the final SuperAlarm patterns, we employ the independent test data set to simulate the application of these SuperAlarm patterns in real-time and assess their performance at predicting code blue events. Based on the method used in [27], at the moment of receiving a new monitor alarm or a new laboratory test result, the algorithm will determine whether any of the final SuperAlarm patterns can be found among the integrated laboratory test results and monitor alarms within a $T_w$-long window preceding the time of this new measurement. It should be noted that $T_w$ is the optimal length of the time window determined in the training process.

By running the simulation across the sequence of monitors and laboratory test results for a given patient, we obtain a new sequence of SuperAlarm triggers. Four metrics are used to assess the performance of SuperAlarm patterns at predicting code blue events:

1. (1) $\text{Sen}^{\alpha}_{\text{TP}}$: sensitivity function with respect to prediction window. This metric is calculated in terms of percentage of code blue patients triggering any of the final SuperAlarm patterns within a prediction window preceding code blue events. This is the same definition used in our previous work.
2. (2) $\text{Sen}^{\beta}_{\text{TP}}$: sensitivity function with respect to lead time. This metric is computed in terms of percentage of code blue patients triggering any of the final SuperAlarm patterns within a time window that starts at 12-th hour and ends at a lead time preceding code blue event.
3. (3) False SuperAlarm ratio. This metric is obtained as a ratio of hourly number of the final SuperAlarm triggers for control patients to that of regular monitor alarms, or that of regular monitor alarms plus laboratory test results if the final SuperAlarm patterns contain laboratory test results.
4. (4) Work-up to detection ratio (WDR). We define the work-up to detection ratio as $\frac{\text{WDR}}{\text{n}}$, where $\text{n}$ is the number of code blue patients triggering any of the final SuperAlarm patterns within a time window preceding code blue events; $\text{b}$ is the number of control patients triggering any of the final SuperAlarm patterns within a window of the same length. The window is randomly selected over the whole monitoring time for each control patient and this process is repeated $M = 1000$ times. Let $T_{ij} = 1(1 \leq i \leq N, 1 \leq j \leq M)$ if any SuperAlarm patterns are triggered within the $j$th selected window in the control patient $i$ and $T_{ij} = 0$ otherwise, where $N$ is the number of control patients in the independent test data set. We estimate the expected value of whether any SuperAlarm patterns are triggered in the control patient $i$ as $\mu_i = \frac{\sum_{j=1}^{M} T_{ij}}{M}$ and the standard deviation of that as $\sigma_i = \sqrt{\sum_{j=1}^{M} (T_{ij} - \mu_i)^2 / M-1}$. The estimated value of $b$ and its standard deviation are then calculated as $\mu_b = \sum_{i=1}^{N} \mu_i$ and $\sigma_b = \sqrt{\sum_{i=1}^{N} \sigma_i^2}$, respectively. At this point, the expected
value and the standard deviation of WDR are finally computed as 
\[ \mu_{WDR} = 1 + \frac{\sum_{i=1}^{N} a_i}{N} \] and 
\[ \sigma_{WDR} = \sqrt{\frac{\sum_{i=1}^{N} a_i}{N^2}}, \]
respectively.

For a given \( FPR_{max} \), we perform the McNemar’s test to determine whether the performances of the three SuperAlarm sets generated from Alarm data set, \( Ab + Alarm \) data set and \( Delta + Lab + Alarm \) data set are significantly different from each other using the independent test data set. To do so, we first partition the data of each control patient into consecutive 4-h windows from the beginning of monitoring to the end. The McNemar’s test is then done as follows. First, we randomly select one of the 4-h windows from each control patient. Next, the three SuperAlarm sets are compared in pairs by applying each of them to both the data of each control patient within the selected 4-h window and the data of each code blue patient within the optimal \( T_{\text{long}} \)-long window preceding code blue events. Third, this process of the McNemar’s test is repeated 1000 times. The performances of any two SuperAlarm sets are considered to be significantly different if the number of significant individual McNemar’s tests is greater than 95% of the total number of tests, which is equivalent to a \( p \)-value of 0.05.

2.6. Patient data

The monitor alarms and laboratory test results in the present study were extracted from a central repository of comprehensive data elements archived for patients hospitalized at the UCLA Ronald Regan Medical Center, Los Angeles, California. Patients involved in this study were from ICUs (neurosurgical, cardiac/thoracic, coronary care, medical, transplant surgical) or other acute care areas (cardiac observation unit, hematology and stem cell transplant unit, medical-surgical specialty unit, nephrology and stroke unit, and liver transplant unit). The Institutional Review Board waiver of consent was obtained for this secondary analysis of the data.

Study subjects include all adult patients (age >18 years) admitted from March 2010 to June 2012 who experienced code blue events. Control patients were admitted within the same period without codes, death, or unplanned ICU transfer. We further refined the selection of control patients by the following criteria [27]:

- Same APR DRG (All Patient Refined Diagnosis Related Group) or Medicare DRG;
- Same age (±5 year);
- Same gender;
- Admission to the same hospital unit within the same month.

254 (54% male) code blue patients with age 61.6 ± 18.2 (mean ± std) and 2213 (68% male) control patients with age 63.5 ± 14.6 were included in this study. Seventy-six percent and 19% of code blue calls were noted for cardiac arrest and respiratory arrest, respectively. Seventy-one percent of code blue patients were admitted in ICUs, 23% in non-ICUs and 6% in other facilities such as operating room and procedure room. On the other hand, 74% of the control patients were from ICUs, 24% from non-ICU units and 2% from other facilities.

3. Results

3.1. Monitor alarms for the code blue patients and control patients

Monitor alarms preceding code blue events were extracted. There were 37 case patients in our data set having more than one code blue call and we only extracted alarms prior to the first code blue call for current analysis. 662,576 raw monitor alarms for code blue patients and 5363019 for control patients were collected. The monitoring time was 250.3 ± 406.1 (mean ± std) hours and 279.9 ± 384.3 h for the case patients and control patients, respectively. Hourly number of monitor alarms was 18.9 ± 27.9 per code blue patient and 9.5 ± 9.8 per control patient. Within a 5 min window preceding code blue event, the number of code blue patients having at least one “crisis” monitor alarm signaling asystole, ventricular fibrillation and no breath was 38 (15.0%), 31 (12.2%) and 3 (1.2%), respectively.

3.2. Laboratory test results for case patients and control patients

We extracted laboratory test results from 19 laboratory test panels, resulting in a total of 62 different laboratory tests. Table 1 provides descriptive statistics of the 19 laboratory test panels. There were 191,483 and 362,960 laboratory results for code blue and control patients, respectively. For code blue patients, 37.1% of laboratory test results were flagged as H while 34.7% as L, 24.9% as N, 2.5% as LL and 0.8% as HH. For control patients, 45.5% of laboratory test results were flagged as H while 41.2% as L, 10.7% as N, 1.8% as LL and 0.8% as HH. It should be noted that the majority of laboratory test results for both case patients and control patients were flagged as either \( H \) or \( L \), indicating that abnormalities were common in the laboratory test results among these patients.

3.3. Results of estimating parameters in Non-Homogenous Poisson Process (NHPP) model

In order to estimate parameters \( \alpha \) and \( \beta \) in the Eq. (1), we first extract monitor alarms for all code blue patients within a 12-h time window preceding code blue events. The 12-h window is divided equally into 24 consecutive subintervals, each 30 min long.

Table 1
Descriptive statistics of the 19 laboratory test panels that are used in the present study.

<table>
<thead>
<tr>
<th>Panel name</th>
<th>% of Results</th>
<th>Code blue patients</th>
<th>% of Patients</th>
<th># of Patients</th>
<th>Control patients</th>
<th>% of Results</th>
<th>% of Patients</th>
<th># of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Ab + Lab )</td>
<td>0.1643</td>
<td>16.643</td>
<td>88.9%</td>
<td>78.091</td>
<td>228</td>
<td>89.8%</td>
<td>7.080</td>
<td>481</td>
</tr>
<tr>
<td>Amylase</td>
<td>0.0224</td>
<td>0.224</td>
<td>10.2%</td>
<td>0.038</td>
<td>26</td>
<td>10.2%</td>
<td>0.039</td>
<td>188</td>
</tr>
<tr>
<td>BNP*</td>
<td>0.2200</td>
<td>2.200</td>
<td>52.0%</td>
<td>0.395</td>
<td>132</td>
<td>52.0%</td>
<td>0.359</td>
<td>175</td>
</tr>
<tr>
<td>CBC*</td>
<td>0.2243</td>
<td>2.243</td>
<td>100.0%</td>
<td>0.257</td>
<td>987</td>
<td>44.6%</td>
<td>0.257</td>
<td>1230</td>
</tr>
<tr>
<td>Chem7</td>
<td>5.7674</td>
<td>57.674</td>
<td>92.9%</td>
<td>5.366</td>
<td>236</td>
<td>36.9%</td>
<td>5.366</td>
<td>236</td>
</tr>
<tr>
<td>Chem10</td>
<td>5.8773</td>
<td>57.877</td>
<td>92.9%</td>
<td>5.766</td>
<td>236</td>
<td>36.9%</td>
<td>5.766</td>
<td>236</td>
</tr>
<tr>
<td>Chem7</td>
<td>3.9093</td>
<td>39.093</td>
<td>54.3%</td>
<td>3.919</td>
<td>138</td>
<td>32.3%</td>
<td>3.919</td>
<td>138</td>
</tr>
<tr>
<td>COAG*</td>
<td>9.9721</td>
<td>9.972</td>
<td>98.8%</td>
<td>10.030</td>
<td>251</td>
<td>40.3%</td>
<td>10.030</td>
<td>251</td>
</tr>
<tr>
<td>CREAT*</td>
<td>6.7322</td>
<td>67.32</td>
<td>88.6%</td>
<td>6.878</td>
<td>225</td>
<td>28.2%</td>
<td>6.878</td>
<td>225</td>
</tr>
<tr>
<td>LIVER*</td>
<td>1.1371</td>
<td>137.1</td>
<td>61.0%</td>
<td>1.759</td>
<td>155</td>
<td>18.8%</td>
<td>1.759</td>
<td>155</td>
</tr>
<tr>
<td>Lipase</td>
<td>0.0117</td>
<td>1.17</td>
<td>6.7%</td>
<td>0.035</td>
<td>17</td>
<td>10.2%</td>
<td>0.035</td>
<td>17</td>
</tr>
<tr>
<td>Liver Func</td>
<td>0.0093</td>
<td>0.93</td>
<td>79.5%</td>
<td>0.004</td>
<td>202</td>
<td>29.7%</td>
<td>0.004</td>
<td>202</td>
</tr>
<tr>
<td>Meds*</td>
<td>0.0560</td>
<td>5.60</td>
<td>7.9%</td>
<td>0.082</td>
<td>20</td>
<td>21.8%</td>
<td>0.082</td>
<td>20</td>
</tr>
<tr>
<td>Troponin</td>
<td>0.8087</td>
<td>80.87</td>
<td>70.5%</td>
<td>0.857</td>
<td>179</td>
<td>26.8%</td>
<td>0.857</td>
<td>179</td>
</tr>
<tr>
<td>Urinalysis</td>
<td>1.9649</td>
<td>19.64</td>
<td>58.7%</td>
<td>2.634</td>
<td>149</td>
<td>29.8%</td>
<td>2.634</td>
<td>149</td>
</tr>
<tr>
<td>CSF*</td>
<td>0.1022</td>
<td>1.02</td>
<td>8.7%</td>
<td>0.079</td>
<td>22</td>
<td>2.6%</td>
<td>0.079</td>
<td>22</td>
</tr>
<tr>
<td>PO2</td>
<td>15.9250</td>
<td>1592</td>
<td>98.4%</td>
<td>14.417</td>
<td>250</td>
<td>40.5%</td>
<td>14.417</td>
<td>250</td>
</tr>
<tr>
<td>Phenobarbitol</td>
<td>0.0201</td>
<td>2.01</td>
<td>3.0%</td>
<td>0.001</td>
<td>3</td>
<td>5.2%</td>
<td>0.001</td>
<td>3</td>
</tr>
<tr>
<td>WBC*</td>
<td>5.2922</td>
<td>52.92</td>
<td>69.7%</td>
<td>4.852</td>
<td>227</td>
<td>7.9%</td>
<td>4.852</td>
<td>227</td>
</tr>
<tr>
<td>Calcium</td>
<td>5.8464</td>
<td>58.46</td>
<td>97.2%</td>
<td>5.186</td>
<td>247</td>
<td>41.3%</td>
<td>5.186</td>
<td>247</td>
</tr>
</tbody>
</table>

\( * \) "# of results" represents percentage of the collected laboratory test results belonging to the given panel. "% of patients" represents the number of code blue patients or controls who have laboratory test results belonging to the given panel.

Please cite this article in press as: Bai Y et al. Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.09.006
After counting the number of alarms in each of the 24 subintervals and applying the GLM model, we obtain the estimated values of the parameters \( \hat{a} = 2.59 \pm 0.20 \) (mean ± std, \( p < 0.01 \)) and \( \hat{b} = -0.08 \pm 0.03 \) (\( p < 0.01 \)).

To set up our experiment, four values of \( T_w \) are assessed: 30 min, 60 min, 90 min and 120 min. Using Eq. (3) with the estimated \( \hat{a} \) and \( \hat{b} \), the value of minimum-alarm-count-threshold for each \( T_w \) is determined as 8 (95% CI: 8.5–19.3), 15 (95% CI: 15.8–38.1), 22 (95% CI: 22.2–56.4) and 27 (95% CI: 27.7–74.4), respectively. Accordingly, the number of excluded code blue patients for each \( T_w \) is 34, 40, 53 and 62, respectively.

### 3.4. Offline analysis results

Three min\_sup thresholds are specified in this study: 0.10, 0.15 and 0.20. This creates 12 algorithm parameter combinations with the four \( T_w \) values. Fig. 3 illustrates the ROC curves for each type of the SuperAlarm sets generated by MAFIA under the various algorithm parameter combinations. We observe that for a given FPR, TPR of the SuperAlarm set generated from the integrated data set is greater than that of the SuperAlarm set from the regular monitor alarm data set.

We specify the acceptable false positive rates \( FPR_{\text{max}} \) as 0.02, 0.05, 0.10 and 0.15 in the present study. Table 2 lists the optimal parameter combinations and the average sensitivity for each type of SuperAlarm sets based on each \( FPR_{\text{max}} \) threshold. It should be noted that this average sensitivity is calculated in the training phase based on the 10-fold CV set. We observe that for a given combination of optimal algorithm parameters, the sensitivity value of a given type of SuperAlarm set grows with increasing \( FPR_{\text{max}} \) threshold. For example, the sensitivity of SuperAlarm set generated from the Delta Lab + Alarm data set increases from 70.7% to 89.0% as \( FPR_{\text{max}} \) increases from 0.02 to 0.15 (11–51, 84–3345, and 59–428 for the Alarm data set, the Ab Lab + Alarm data set, and the Delta Lab + Alarm data set, respectively).

An example of the SuperAlarm pattern generated from the Delta Lab + Alarm data set is given as follows: SuperAlarm pattern “BRADY; APTT H ? H; WBC H ? H; Pt H ? H; Ca, plasma L ? L; Hematocrit L ? L; Hemoglobin L ? L”, which represents that if a patient was bradycardia, and the activated partial thromboplastin time (APTT), white blood cell count (WBC), prothrombin time (PT) remained high, but plasma calcium, hematocrit and hemoglobin remained low, then the patient may be at high risk.

### 3.5. Simulated online analysis results

Fig. 4 shows the curves of Sen\_P@T based on the four \( FPR_{\text{max}} \) thresholds. We also plot the sensitivities of regular monitor alarms with and without “crisis” alarms, respectively. The sensitivity of

<table>
<thead>
<tr>
<th>( FPR_{\text{max}} )</th>
<th>SuperAlarm type</th>
<th>( Optimal \ T_w ) (min)</th>
<th>( Optimal \ min_sup )</th>
<th>Sensitivity% (mean ± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>Alarm</td>
<td>30</td>
<td>0.10</td>
<td>48.6 ± 15.4</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>120</td>
<td>0.10</td>
<td>58.0 ± 13.7</td>
</tr>
<tr>
<td></td>
<td>Delta lab + Alarm</td>
<td>30</td>
<td>0.10</td>
<td>70.7 ± 18.1</td>
</tr>
<tr>
<td>0.05</td>
<td>Alarm</td>
<td>60</td>
<td>0.10</td>
<td>62.6 ± 14.1</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>60</td>
<td>0.15</td>
<td>70.6 ± 14.3</td>
</tr>
<tr>
<td></td>
<td>Delta lab + alarm</td>
<td>30</td>
<td>0.10</td>
<td>76.2 ± 18.9</td>
</tr>
<tr>
<td>0.10</td>
<td>Alarm</td>
<td>60</td>
<td>0.10</td>
<td>70.9 ± 12.6</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>60</td>
<td>0.20</td>
<td>80.8 ± 15.9</td>
</tr>
<tr>
<td></td>
<td>Delta lab + Alarm</td>
<td>90</td>
<td>0.15</td>
<td>83.3 ± 12.2</td>
</tr>
<tr>
<td>0.15</td>
<td>Alarm</td>
<td>60</td>
<td>0.10</td>
<td>79.6 ± 13.1</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>90</td>
<td>0.10</td>
<td>85.0 ± 9.7</td>
</tr>
<tr>
<td></td>
<td>Delta lab + Alarm</td>
<td>30</td>
<td>0.10</td>
<td>89.0 ± 10.3</td>
</tr>
</tbody>
</table>

Fig. 3. Receiver operator characteristic (ROC) curves of the three SuperAlarm sets generated under different combination of algorithm parameters in the offline training phase. The row represents min\_sup thresholds while the column represents \( T_w \)-long time window (min). Since ROC curve with given min\_sup and \( T_w \) is obtained by averaging the 10 ROC curves generated from the 10-fold CV set, we additionally mark maximum standard deviation for each of ROC curves using error bar.

Please cite this article in press as: Bai Y et al. Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.09.006
regular monitor alarms with respect to prediction window is calculated in terms of percentage of code blue patients having any regular monitor alarms within the prediction window. As we expected, the sensitivity of regular monitor alarms with “crisis” alarms is greater than that without “crisis” alarms when time is near the onset of code blue events. We can observe that for a given $FPR_{\text{max}}$, the Sen$_{P\@T}$ value of a SuperAlarm set becomes higher as the length of prediction window is extended. We also observe that for a given length of prediction window, the Sen$_{P\@T}$ value of a SuperAlarm set increases with $FPR_{\text{max}}$ threshold (specificity decreases).

Fig. 5 displays the curves of Sen$_{L\@T}$ based on the four $FPR_{\text{max}}$ thresholds. Sensitivities for regular monitor alarms with respect to lead time with and without “crisis” alarms are also shown. Here the sensitivity for regular monitor alarms is calculated as percentage of code blue patients having any of regular monitor alarms within 12-h window prior to the lead time. We observe that no matter what the lead time is, the sensitivity for regular monitor alarms when time is near the onset of code blue events. We can observe that for a given $FPR_{\text{max}}$, the Sen$_{L\@T}$ value of a SuperAlarm set becomes higher as the length of prediction window is extended. We also observe that for a given length of prediction window, the Sen$_{L\@T}$ value of a SuperAlarm set increases with $FPR_{\text{max}}$ threshold (specificity decreases).

In addition, Table 3 lists the false SuperAlarm ratio and the work-up to detection ratio for each type of SuperAlarm set based on a given $FPR_{\text{max}}$ threshold. We also report the sensitivities with respect to different lengths of prediction window and lead time of half hour, 1 h, 2 h, 6 h and 12 h, respectively. From these results, we can see that for a given type of SuperAlarm set, a higher $FPR_{\text{max}}$ threshold leads to a higher Sen$_{P\@T}$ and a higher Sen$_{L\@T}$ but also a larger work-up to detection ratio and a larger false SuperAlarm ratio. Taken as an example the SuperAlarm set generated from the Delta Lab + Alarm data set when the $FPR_{\text{max}}$ threshold increases from 0.02 to 0.15, the Sen$_{P\@T}$ value for 1-h prediction window and the Sen$_{L\@T}$ value for 1-h lead time increase from 56.7% to 93.3% and from 43.3% to 90.0%, respectively. However, the false SuperAlarm ratio and the work-up to detection ratio within 12-h window also rise from 2.0% to 14.8% and from 2.1 to 6.5, respectively. We can also observe that when $FPR_{\text{max}} = 0.15$, for instance, the Sen$_{L\@T}$ value of the SuperAlarm set generated from the Delta Lab + Alarm data set reduces from 93.3% to 80.0% with the extension of the length of lead time from half hour to 12 h. It can be seen that for a given $FPR_{\text{max}}$ threshold and a given length of window, the Sen$_{P\@T}$ and the Sen$_{L\@T}$ of the SuperAlarm set generated from the Delta Lab + Alarm or the Ab Lab + Alarm data set are higher than that of the SuperAlarm set generated from the Alarm data set, whereas the work-up to detection ratio of SuperAlarm set generated from the Delta Lab + Alarm or the Ab Lab + Alarm data set...
Fig. 5. Sensitivity curves of the three final SuperAlarm sets with respect to lead time (i.e., SenL@T) based on $FPR_{max} = 0.02$ (A), $FPR_{max} = 0.05$ (B), $FPR_{max} = 0.10$ (C) and $FPR_{max} = 0.15$ (D), respectively. The x-axis represents the length of lead time preceding code blue events. The magenta curve and black curve represent the sensitivity of regular monitor alarms with respect to lead time with and without “crisis” alarms from the independent test data set, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Performance metrics of the sensitivity with respect to prediction window (SenP@T), the sensitivity with respect to lead time (SenL@T), the false SuperAlarm ratio and the work-up to detection ratio. These metrics are calculated by applying the final SuperAlarm set to the independent test data set based on varying $FPR_{max}$ thresholds.

<table>
<thead>
<tr>
<th>$FPR_{max}$</th>
<th>SuperAlarm type</th>
<th>Sensitivity (%)</th>
<th>False SuperAlarm ratio (%; mean ± std)</th>
<th>Work-up to detection ratio (mean ± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metrics</td>
<td>Half hour</td>
<td>1 h</td>
<td>2 h</td>
</tr>
<tr>
<td>0.02</td>
<td>Alarm</td>
<td>SenP@T</td>
<td>36.7</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>40.0</td>
<td>36.7</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>SenP@T</td>
<td>40.0</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>40.0</td>
<td>36.7</td>
</tr>
<tr>
<td>0.05</td>
<td>Alarm</td>
<td>SenP@T</td>
<td>43.3</td>
<td>43.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>43.3</td>
<td>33.3</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>SenP@T</td>
<td>70.0</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>63.3</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td>Delta lab + Alarm</td>
<td>SenP@T</td>
<td>66.7</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>76.7</td>
<td>73.3</td>
</tr>
<tr>
<td>0.10</td>
<td>Alarm</td>
<td>SenP@T</td>
<td>53.3</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>53.3</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>SenP@T</td>
<td>80.0</td>
<td>80.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>76.7</td>
<td>73.3</td>
</tr>
<tr>
<td></td>
<td>Delta lab + Alarm</td>
<td>SenP@T</td>
<td>76.7</td>
<td>76.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>83.3</td>
<td>80.0</td>
</tr>
<tr>
<td>0.15</td>
<td>Alarm</td>
<td>SenP@T</td>
<td>66.7</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>86.7</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>Ab lab + Alarm</td>
<td>SenP@T</td>
<td>83.3</td>
<td>83.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SenL@T</td>
<td>86.7</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>Delta lab + Alarm</td>
<td>SenP@T</td>
<td>93.3</td>
<td>93.3</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Bai Y et al. Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.09.006
set is smaller than that of the SuperAlarm set generated from the Alarm data set.

Fig. 6 shows SuperAlarm triggers within a 12-h window preceding code blue events from the independent test data set consisting of 30 code blue patients. In each row of the plot, a white dot is placed at the time of a SuperAlarm trigger. These SuperAlarm triggers are from the largest SuperAlarm set obtained under the FPRmax threshold of 0.15 from the Alarm data set (Fig. 6(A)), the Ab Lab + Alarm data set (Fig. 6(B)) and the Delta Lab + Alarm data set (Fig. 6(C)), respectively. It can be seen that the SuperAlarm triggers become more frequent as time approaches the onset of code blue events. We also observe that the SuperAlarm triggers generated from the Delta Lab + Alarm data set or the Ab Lab + Alarm data set are more frequent than that generated from the Alarm data set. These visual assessments match the quantitative results reported above.

For the FPRmax threshold of 0.02, a large majority (904 and 983) of the 1000 repeated McNemar’s tests, conducted on randomly selected data, shows that the performances of SuperAlarm sets generated from the Ab Lab + Alarm data set and from the Delta Lab + Alarm data set are significantly different from that of the SuperAlarm set generated from the Alarm data set, respectively. Only 117 (11.7%) tests show that performance of SuperAlarm set generated from the Ab Lab + Alarm data set is significantly different from that of SuperAlarm set generated from the Delta Lab + Alarm data set. These McNemar’s tests demonstrate that the performances of the SuperAlarm sets generated from the Ab Lab + Alarm data set and from the Delta Lab + Alarm data set under the optimal algorithm parameters are significantly different from that of the SuperAlarm set generated from the Alarm data set. However, the performance of the SuperAlarm set from the Ab Lab + Alarm data set is not significantly different from that of SuperAlarm set from the Delta Lab + Alarm data set. For the FPRmax thresholds of 0.05, 0.10 and 0.15, we can draw the same conclusion because the numbers of the McNemar’s tests resulting in significantly different performances between these three types of SuperAlarm sets are [982 (98.2%), 979 (97.9%) and 248 (24.8%)], [967 (96.7%), 984 (98.4%) and 304 (30.4%)], and [962 (96.2%), 974 (97.4%) and 171 (17.1%)], respectively.

4. Discussion

In this study we have detailed the approaches and results from advancing a methodological framework of utilizing patient monitor alarms and laboratory test results to detect patient deterioration. Several new algorithmic elements have been introduced to the SuperAlarm framework that was created in our previous work [27]. Specifically, we excluded “crisis” alarms and code blue patients with unexpectedly small number of monitor alarms. We proposed two approaches to integrate monitor alarms with laboratory test results. The SuperAlarm patterns were discovered using MAFIA, which produced less redundant SuperAlarm patterns than those produced by Apriori-based methods used in our previous work. The results based on an independent test data set showed that SuperAlarm patterns discovered from the integrated data set of monitor alarms along with laboratory test results achieved higher sensitivity to predict code blue events and have fewer false triggers for control patients.

Patients studied here might have abnormally small number of monitor alarms for two broad reasons. It may be due to technical reasons such as data loss from the data acquisition system or due to pathophysiological reasons that sudden patient deterioration was not preceded by many alarms. If we included cases with small number of alarms due to technical reasons, the SuperAlarm’s sensitivity would thereby be incorrectly estimated. Hence, we estimated the most likely minimum count of monitor alarms for code blue patients (i.e., minimum-alarm-count-threshold) and excluded those code blue patients whose count of alarms were less than the minimum-alarm-count-threshold. This practice may have excluded cases with small number of alarms due to pathophysiological reasons. Since this study was retrospective, we were not able to differentiate these two causes for a given case. Nevertheless, adopting the NHPP model to exclude patients may not impact the validity of our results for two reasons. First, it is known that ventricular fibrillation (VFib) cardiac arrest can occur suddenly. We therefore checked the number of excluded patients with VFib alarms. However, for each of the Tw-long time windows assessed in the present study, only 1 out of 34, 2 out of 40, 2 out 53 and 4 out of 62 excluded code blue patients had VFib alarms, respectively. Second, we did not exclude patients from the independent test data set and therefore the reported sensitivity may actually be an underestimate of its true value considering that patients with small number of alarms due to data loss may be included.

Compared to the SuperAlarm set consisting of only monitor alarms, the SuperAlarm set composed of monitor alarms and laboratory test results achieved higher sensitivity and lower work-up to detection ratio under an acceptable false positive rate (FPRmax). As we reported in Section 3.4, both SuperAlarm sets generated from Ab Lab + Alarm data set and Delta Lab + Alarm data set yielded better performance than that generated from Alarm data set in terms of sensitivity to predict code blue events and the value of work-up to detection ratio. One likely explanation for this better performance might be that the laboratory test results provided more information about the patient’s condition. Another reason, according to Table 1, may be related to the fact that code blue patients on average had more laboratory tests performed, reflecting a higher clinical demand of those laboratory tests to manage patients whose clinical status were declining.

In the present study, both abnormal laboratory test results and delta laboratory test results were represented based on whether they were out of the standard range reference. With this representation we were able to simplify the process of integrating...
monitor alarms because both data modalities can now be treated as discrete events. However, this representation did not take into account the numeric values of the laboratory test results. Escobar et al. [37] developed a model to predict non-ICU patient deterioration where a laboratory-based acute physiology score (LAPS) based on numeric values from 14 laboratory test results was used. Their study suggested that SuperAlarm might be improved by further considering ways to incorporate numeric values of laboratory test results. In addition, although 62 laboratory tests from 19 laboratory panels were integrated with monitor alarms here to build SuperAlarm set, only a subset (up to 35) of these 62 laboratory test results were part of the SuperAlarm patterns. On the other hand, there would be other laboratory tests that might be highly correlated with patient deterioration. Future work would also focus on investigation into whether different laboratory tests would improve the performance at predicting deterioration. Apart from integration of laboratory tests with monitor alarms, there is still a great volume of relevant data within an Electronic Medical Record (EMR) system that can be used to predict patient deterioration. Heldt et al. [38] suggested that an advanced patient monitor alarms because both data modalities can now be treated as discrete events. However, this representation did not take into account the numeric values of the laboratory test results. Escobar et al. [37] developed a model to predict non-ICU patient deterioration where a laboratory-based acute physiology score (LAPS) based on numeric values from 14 laboratory test results was used. Their study suggested that SuperAlarm might be improved by further considering ways to incorporate numeric values of laboratory test results. In addition, although 62 laboratory tests from 19 laboratory panels were integrated with monitor alarms here to build SuperAlarm set, only a subset (up to 35) of these 62 laboratory test results were part of the SuperAlarm patterns. On the other hand, there would be other laboratory tests that might be highly correlated with patient deterioration. Future work would also focus on investigation into whether different laboratory tests would improve the performance at predicting deterioration. Apart from integration of laboratory tests with monitor alarms, there is still a great volume of relevant data within an Electronic Medical Record (EMR) system that can be used to predict patient deterioration. Heldt et al. [38] suggested that an advanced patient monitoring system should integrate and analyze multi-dimensional clinical variables including alarms, waveforms, vital signs, laboratory tests and clinical notes to monitor the pathophysiological state of a patient. Huang et al. [39] reported that surveillance tools in modern hospitals may benefit from the integration of early warning scores with medications that are temporally associated with clinical deterioration to improve patient outcomes. By design, SuperAlarm is inherently a multivariate approach designed to recognize patient deterioration. Therefore, it meets the requirement of a patient-centered design of future patient alarm systems which should integrate patient data and assess clinical patterns of multiple alarms and associated vital signs holistically [40].

We employed the MAFIA algorithm to generate the SuperAlarm patterns in the present work. MAFIA was designed to discover patterns with maximal number of components that still satisfy the minimum support threshold [35]. This is a desirable characteristic because a long SuperAlarm pattern is less likely to be triggered by control patients. In our previous work [27], we extracted frequent itemsets (FI) and closed frequent itemsets (CFI) as SuperAlarm patterns, which may likely contain redundant SuperAlarm patterns leaving room for more frequent false triggers. However, the current algorithm will not recognize potentially useful patterns embedded in the occurring order of alarms and laboratory tests. Additional approaches such as Hidden Markov Model (HMM) [41], Bayesian Network [42] and String kernels [43] should be investigated as potential methodological improvement to the SuperAlarm framework.

In this study we were not able to implement other algorithms of detecting patient deterioration or compared their performance with that of SuperAlarm. This is partly due to the fact that our existing data set does not contain vital sign data or nursing notes that are needed in several existing patient deterioration detection algorithms [20,23,24,37]. Nevertheless, we presented here the performance of these algorithms as reported in the original papers. In [20], the authors demonstrated that MEWS score \( \geq 5 \) was associated with increased risk of death, ICU admission and high dependency unit (HDC) admission with odds ratio being 5.4, 10.9 and 3.3, respectively. In [23], the authors reported that the positive predictive value (PPV) of the Biosig alert was 95%. In [24], the authors reported that RI predicted patient deterioration, 24-h mortality, and 30-day readmissions with a c-statistics > 0.92, > 0.93 and >0.62, respectively. In [37], Escobar et al. reported their model predicted patient deterioration outside the ICU with a c-statistic value of 0.775 in the validation dataset and the work-up to detection ratio was 14.5 when identifying 15% of all transfers to the ICU. It is important to select appropriate performance metrics to help users evaluate patient deterioration detection systems. Conventional metrics as c-statistics undoubtedly have strong theoretical underpinnings. However, we argue that a patient deterioration detection system needs to be evaluated at a particular operating point on the ROC curve. At a chosen operation point, work-up to detection ratio is an excellent metric to gauge the extra work for a correct detection and can be readily communicated to clinical users. In addition, sensitivity can help understand how many deterioration events can be potentially captured. For monitoring applications, sensitivity needs to be evaluated as a function of lead time. However, the concept of incorporating lead time in evaluating sensitivity has not been widely used. To better compare with monitor alarm frequency, the false SuperAlarm ratio is proposed. This metric has not been used in other works either. In summary, we acknowledge that a direct comparison of similar patient deterioration detection approaches needs to be done, preferably using a standard database, proper implementation, and appropriate performance metrics.

Finally, the discovered SuperAlarm patterns need further verifications by clinical knowledge. Given that these patterns were discovered from data of critically ill patients, it is very likely that some of these patterns may not add new knowledge per se but being able to track these patterns automatically in practice may alleviate the alarm fatigue problem.

5. Conclusion

The present study proposed novel approaches to integrate monitor alarms with laboratory test results to discover SuperAlarm patterns using maximal frequent itemsets mining technique. The performance of SuperAlarm patterns were assessed based on four metrics using an independent test data set: sensitivity with respect to prediction window, sensitivity with respect to lead time, false SuperAlarm ratio, and work-up to detection ratio. Results showed that both the SuperAlarm sets generated from \( \text{Ab lab} + \text{Alarm} \) data set and \( \Delta \text{lab} + \text{Alarm} \) data set outperformed the SuperAlarm set consisting of only monitor alarms in terms of these metrics. Further performance gain may be achieved by using numeric values of laboratory test results, integrating metrics of raw physiological signals as additional “alarms”, and incorporating sequential patterns of SuperAlarm triggers.

References


