HHS Public Access

Author manuscript

ACM BCB. Author manuscript; available in PMC 2016 August 14.

Published in final edited form as:

ACM BCB. 2014 September ; 2014: 211–219. doi:10.1145/2649387.2649440.

icuARM-II: improving the reliability of personalized risk prediction in pediatric intensive care units

Chih-Wen Cheng, **Nikhil Chanani**, **Kevin Maher**, and **Wang, M.D. [IEEE Senior Member]**

Abstract

Clinicians in intensive care units (ICUs) rely on standardized scores as risk prediction models to predict a patient's vulnerability to life-threatening events. Conventional Current scales calculate scores from a fixed set of conditions collected within a specific time window. However, modern monitoring technologies generate complex, temporal, and multimodal patient data that conventional prediction models scales cannot fully utilize. Thus, a more sophisticated model is needed to tailor individual characteristics and incorporate multiple temporal modalities for a personalized risk prediction. Furthermore, most scales models focus on adult patients. To address this needdeficiency, we propose a newly designed ICU risk prediction system, called icuARM-II, using a large-scaled pediatric ICU database from Children's Healthcare of Atlanta. This novel database contains clinical data collected in 5,739 ICU visits from 4,975 patients. We propose a temporal association rule mining framework giving clinicians a potential to perform predict risks prediction based on all available patient conditions without being restricted by a fixed observation window. We also develop a new metric that can rigidly assesses the reliability of all all generated association rules. In addition, the icuARM-II features an interactive user interface. Using the icuARM-II, our results demonstrated showed a use case of short-term mortality prediction using lab testing results, which demonstrated a potential new solution for reliable ICU risk prediction using personalized clinical data in a previously neglected population.

1. Introduction

Critically ill patients need intensive care due to impaired vital functions that have to be monitored with a higher frequency and fidelity, and at greater resource utilization. One of the main tasks in intensive care is to provide continuous treatments until the patient's body recovers to resume these functions. The modern intensive care unit (ICU), equipped with high-level body sensing and monitoring, generates a large volume of complex and multimodal data that allows clinicians to provide timely and effective treatments. However, because of the limitation of human intellectual abilities [1], making sense of such comprehensive but heterogeneous data with hundreds of variables becomes impossible. In other words, the data becomes richer, but the extracted knowledge is still limited, which raises the need for computer-based data mining techniques to assist in patient care.

*Corresponding Author: maywang@bme.gatech.edu, Phone: 404-385-2954, Fax: 404-894-4243, Address: Suite 4106, UA Whitaker Building, 313 Ferst Drive, Atlanta, GA 30332, USA.

Developing risk prediction models is one of the major purposes of ICU data mining [2]. Models with significant validations and refinements became widely used illness scoring systems, such as Acute Physiology and Chronic Health Evaluation (APACHE) [3], Mortality Prediction Model (MPM) [4], Simplified Acute Physiology Score (SAPS) [5], Multiple Organ Dysfunction Score (MODS) [6], Sequential Organ Failure Assessment (SOFA) [7], and Logistic Organ Dysfunction Score (LODS) [8]. More recent studies applied advanced data mining approaches to develop risk prediction models that can handle more complicated clinical situations. For example, authors in [9] utilized fuzzy modeling and tree search feature selection to predict ICU readmissions; another study in [10] applied a combination of statistical models to predict prolonged mechanical ventilation. However, a majority of these risk prediction models share two common limitations—fixed attributes and fixed observation periods.

The development of the conventional risk prediction model started with a target clinical problem (e.g., mortality or prolonged ICU stay). Then researchers selected a set of attributes and applied feature selection methods to extract determinant ones. Finally, researchers applied machine-learning techniques to construct prediction models followed by appropriate validations. The goal of such process is to use as few attributes as possible to achieve a high prediction accuracy. However, a model with fixed variables may be applicable only to some "global" conditions (e.g., heart rate or blood pressure). It is challenging to adopt such models for describing individual characteristics that are outside the model's conditions. Therefore, even though the model can provide evidence, the prediction still needs to be subjectively adjusted with many out-of-scope/extra conditions. Such a process highly relies on a clinician's knowledge and experience, which introduces uncertainty and human biases in the final decision [11].

In addition to the issue of fixed attributes, conventional prediction models used values acquired in a fixed time period. For instance, Zygun and others have used data in the first 24 hours after admission to predict ICU mortality [12–14]. However, models with fixed observation periods may ignore the progressive nature of patient's conditions. For example, a patient's glucose level on ICU day three is likely to be different from that on the admission day but potentially just as relevant. Even though several studies provide prediction models for days other than the day of admission, no clear discriminations were found in comparison to those only on the admission day [15]. In addition, many models consider only the most abnormal values, which means a patient with five times an abnormal level is treated the same as another patient with once mildly elevated level.

Based on the two aforementioned limitations, there is a need to investigate and develop risk prediction models that are able to incorporate all possible conditions from a patient without being constrained by a specific observation period. In this study, we present an ICU risk prediction system, called icuARM-II, based on our previous study in [16]. We propose a temporal association rule mining framework that allows clinicians to construct prediction models based on a patient's personalized conditions in an arbitrary observation period. We structure the remainder of this paper as follows. In Section 2, we introduce the temporal association rule mining framework with a validation process. In Section 3, we provide a case study using lab testing for the prediction of short-term ICU mortality to demonstrate its

usability. In Section 4, we describe icuARM-II's user interface. The conclusion and future directions are summarized in Section 5.

2. TEMPORAL ASSOCIATION RULES

2.1 Principle of association rule mining

Association rule mining (ARM) is one of main data mining areas other than classification and clustering [17]. ARM enables the discovery of all possible relationships among input variables. An association rule is in the form of $A \Rightarrow C$, implying that, if an antecedent A occurs, a consequent C may also likely occur. The A and C are itemsets that are composed of one or multiple items and are mutually exclusive. Agrawal *et al.* first proposed ARM for the purpose of market basket analysis [18] in which a rule $A \Rightarrow C$ carries the implication that if a customer purchases items in A, he/she is also likely to purchase items in C.

The quality of an association rule is quantified by two important metrics—*support* and *confidence*. The support is defined as the fraction of data tuples in the dataset that contain all items in both A and C. For example, an association rule with support of 75% indicates that 75% of the data contains both the antecedent A and the consequent C of the rule. A rule with high support indicates that the rule is frequent in the mining dataset. The second metric of an association rule is called confidence. It determines the possibility of the occurrence of C given A. For instance, a rule with 95% confidence implies that 95% of the tuples that contain A also contain C. A rule with high confidence suggests a strong association between A and C. In order to ensure the quality of mined rules, the mining process requires a minimum support ($Supp_{min}$) and a minimum confidence ($Conf_{min}$) to prune infrequent and unconfident rules. Interested readers can refer to [19] for more detail regarding the generation of frequent itemsets and confident rules.

2.2 Applications of association rule mining

ARM has been widely used in data mining for the discovery of new and potentially important relationships among variables in biomedical and healthcare domains such as gene expression profiling [20, 21], cardiovascular disease prediction [22, 23], healthcare auditing [24, 25], neurological diagnosis [26, 27], and predictive health [28]. In our previous study, we developed an ICU clinical decision support system (named *icuARM*) using ARM and MIMIC-II, which is a well-known ICU research database [29]. The icuARM system not only demonstrated the possibility of personalized data mining in the ICU but also provided an interactive user interface for real-time clinical decision support. However, a majority of ARM applications only consider co-occurrence between antecedent and consequent without the ability to reveal their temporal relationships. For example, a rule {*blocked urinary tract*} $\Rightarrow$ {*creatinine* > *1.3 mg/dL*} can only be used to find the coexistence between *blocked urinary tract* and *creatinine* > *1.3 mg/dL* at the same time. It is not capable of revealing if a patient who has blocked urinary tract will develop abnormal creatinine level in the next 24 hours, even though the patient's creatinine level is currently normal. Although a few temporal ARM approaches have been proposed [30, 31], they were not designed for flexible time window in the antecedent and consequent, and they lacked applications in the healthcare data mining. Therefore, in this study we introduce a temporal association rule

mining framework that can flexibly capture the nature of temporal relationships between the antecedent and the consequent.

2.3 Temporal association rules

Similar to the items in non-temporal association rules, the basic element in temporal association rules is called an *event*. Given a temporal database with time-stamped entities, we use two types of mechanisms to generate event sequences:

- *state events*: which convert qualitative values into categorical (e.g., low, high, or normal) states in a time series, and
- *trend events*: which capture continuous trend courses (e.g., increasing, decreasing, or stationary) in a time series.

Assuming a set of events $E = \{E_1, E_2, \ldots, E_N\}$ has already been defined, we can construct an *episode* $P = \{E_P | T_P\}$ with a set of events $E_P \subseteq E$ within a time interval T_P. T_P is composed of $T_{P,s}$ and $T_{P,e}$, $T_{P,e} > T_{P,s}$, indicating the start time and the end time of T_P. $T_{P,s}$ and $T_{P,e}$ can be specified, for example, $T_P = \{T_{admission}, T_{admission} + 24\text{-}hr\}$ specifies a time interval between the time of admission and 24 hours after then. T_P can also be an arbitrary interval; for instance, $|T_P| < 12\text{-}hr$ means all $T_{P,s}$ and $T_{P,e}$ that $T_{P,e} - T_{P,s} < 12\text{-}hr$. In this way, a temporal association rule (TAR) can be represented as $A \Rightarrow C$: $\{E_A | T_A\} \Rightarrow \{E_C | T_C\}$. All of the E_A, E_C, T_A, and T_C are the user inputs for the mining process. The rule indicates that when an antecedent episode A with events E_A observed within the past T_A, another consequent episode C with E_C in the following T_C will also be likely to occur in a certain possibility. We restrict the mining to the case where an antecedent episode is followed by a consequent episode, i.e., $T_{A,e} = T_{C,s}$. Several use cases can be derived based on different settings of T_A and T_C, as listed and illustrated in Table 1.

2.4 Mining temporal association rules

Given all available event sequences and target antecedent and consequent episodes, our framework counts a rule using a 2×2 contingency table that can be used to derive a variety of rule metrics, including the support and confidence. As shown in Figure 1, the contingency table of a rule $A \Rightarrow C$: $\{E_A | T_A\} \Rightarrow \{E_C | T_C\}$ is presented by four cells, c_{11}, c_{12}, c_{21}, and c_{22}, which are the counts of (A, C), $(A, \overline{C})$, $(\overline{A}, C)$, and $(\overline{A}, \overline{C})$, respectively. $\overline{A}$ (or $\overline{C}$) indicates the situation that the antecedent A (or the consequent C) is not detected. If the start time and end time of T_A (or T_C) are specified (e.g., $T_A = \{T_{admission}, T_{admission} + 12\ hr\}$), the framework directly extracts all events within that time window and determines if the extracted events match E_A (or E_C). Based on the result, the framework updates the four cells accordingly. If the time window of T_A (or T_C) is arbitrary (e.g., $|T_A| < 12\text{-}hr$), then the framework performs two types of scanning on event sequences to count these four cells.

2.4.1 Backward and forward scanning—If T_A and T_C are arbitrary, the framework performs a backward scanning followed by a forward scanning on each event sequence, as depicted in Figure 2. For an event sequence, the backward scanning starts from the last event and traces back towards the first event. Whenever an antecedent episode A is found (i.e., scanned events contain E_A within a time window T_A) with the last event occurring at time t,

the framework extracts potential consequent events between t and $t + T_C$. If potential consequent events contain E_C (i.e., C is scanned), then the framework adds c_{11} by one, and the scanning stops for this event sequence; otherwise, the backward scanning continues. If the backward scanning can only detect an antecedent episode but no consequent episode throughout the sequence (i.e., $(A, \overline{C})$), the framework proceeds to the forward scanning.

The forward scanning starts from the first event towards the last event of a sequence. Whenever a consequent episode C is found (i.e., events contain E_C within a time window T_C) with the first event occurring at t, the process extracts all possible antecedent events occur between $t - T_A$ and t. If all extracted antecedent events contain E_A, the process updates the contingency table depending on whether $(A, \overline{C})$ has been detected in the backward scanning phase. If yes, the framework adds both c_{12} and c_{21} by half; otherwise, it adds only c_{21} by one. If the process finds no matched episodes for E_C in forward scanning, again, the process updates the contingency table depending on whether $(A,)$ has been detected in the backward scanning phase. If yes, the framework adds c_{12} by one; otherwise, it adds c_{22} by one.

Our scanning mechanism ensures that the sum of the four cells in the contingency table is added one by one for each event sequence. It is possible that both c_{21} and c_{21} are added by half since a sequence can have both antecedent episode and consequent episode occur separately, as the example b in Figure 2.

2.4.2 Classification-based rule generation—Our proposed framework assumes a classification-based rule generation from which all rules share one and only one pre-determined consequent episode C (i.e., the class). Let $E = \{E_1, E_2, \ldots, E_N\}$ are N possible antecedent events (i.e., observed patient conditions), the framework aims to discover a *decision list* in which are all confident temporal association rules with different combinations of these events in antecedents. The framework utilizes a two-step process to generate frequent and confident rules according to the specified $Supp_{min}$ and $Conf_{min}$, respectively. The first step is iterative, starting by generating contingency tables of candidate 1-event rules that contain only one event in the antecedent episode. Assuming the total number of event sequences is N_S, candidate 1-event rules that have c_{11}/N_S lower than $Supp_{min}$ are pruned out and the remaining ones are called frequent 1-event rules. In the following iterations (i.e., k>1), the framework first uses frequent (k-1)-event rules to generate candidate k-event rules. Candidate k-event rules that have c_{11}/N_S lower than $Supp_{min}$ are pruned out and the remaining ones are called frequent k-event rules. The iteration continues until no more frequent rules can be found. Given all frequent rules, in the second step, the framework prunes out rules that have $c_{11}/(c_{11}+c_{12})$ lower than $Conf_{min}$, and the remaining ones are called confident rules. In this way, N potential antecedent events can generate a decision list with up to $(2^N - 1)$ raw rules, and infrequent or unconfident rules can be pruned by applying the $Supp_{min}$ and $Conf_{min}$, respectively.

2.5 Reliability metric

2.5.1 Reliability assessment using the principle of leave-one-out cross validation—Given a set of observed patient conditions as antecedent events, the rule

generation phase can end up with a decision list with rules. Afterwards, we apply the principle of leave-one-out (LOO) to evaluate the overall reliability of these rules. As illustrated in Figure 3, given a dataset with event sequences from N_S patients, the LOO process generates the decision list using the event sequences from $N_S - 1$ patients and counts how many number of rules' antecedents that the remaining patient sequence verifies. If the percentage of verified antecedents exceeds a specified class threshold (TH), the sequence satisfies the consequent class. After N_S iterations of the LOO validation, we can obtain the sensitivity and specificity based on the current TH. Then a receiver operating characteristic (ROC) curve can be constructed by changing the TH from 0% to 100%. Finally, the reliability of all rules in the decision list is determined by the corresponding area under the ROC curve (AUC), which is more preferably than considering accuracy alone [32].

2.5.2 Creation, storage, and retrieval of performance—As illustrated in Figure 3, after the rule generation and validation phases, the framework constructs a performance data for the target antecedent and consequent episodes. A performance data consists of components including (1) target antecedent episode A, (2) targeted consequent episode C, (3) the decision list with all generated rules, (4) the ROC curve, and (5) the reliability (i.e., AUC). Because the rule generation and validation require a large amount of computation resources and time, it is necessary to store all generated performance data to avoid duplicate computation. As depicted in Figure 4, upon receiving a new antecedent episode and consequent episode, the framework searches the performance database for existing performance data that contains the input episodes. If found, the framework extracts the rule with the highest confidence and displays the ROC curve with the corresponding reliability. If not found, the framework performs rule generation and validation and updates the database with the new performance data.

2.5.3 Reliability metric vs. classification based on association—Instead of support and confidence that are metrics of individual rules, our reliability metric is proposed to evaluate the quality of a decision list with all generated rules. One of the associative classification algorithms is Classification Based on Association (CBA), which was empirically found to be more accurate than C4.5 on a variety number of datasets [33]. CBA uses a heuristic method to construct a classifier in which all rules in the decision list are ordered decreasingly according to their confidence and support. When classifying a new tuple, the first rule satisfying the tuple is used to classify it. For example, a set of antecedent events $E = \{E_1, E_2, E_3, E_4\}$ can generate a decision list of rules in which the first rule R_1 with antecedent events $\{E_1, E_2\}$ has a high confidence of 99%. The CBA classifies a new tuple according to this rule. However, the decision list may contain other two rules R_2 and R_3 with antecedent events $\{E_1, E_3, E_4\}$ and $\{E_1, E_4\}$ with confidence values of 40% and 38%, respectively. If all rules in the decision list are similar to R_2 and R_3 that tend to have low confidence values, it may imply that the decision list generated is not reliable because not all generated rules can guarantee not only high but also consistent confidence values. The high confidence of R_1 might happen by chance. In this situation, classifying a new tuple using R_1 in CBA becomes problematic. Therefore, after generating a decision list with a set of rule, the use of our classification approach was expected to improve the performance

since it classifies a tuple according to the "voting-based" percentage of verified antecedents in the rules of a decision list, instead of only the top rule with the highest confidence.

3. RESULTS AND DISCUSSION

3.1 CHOA ICU database

After being approved by the Institutional Review Board (IRB), we imported the data in icuARM-II from the Children's Healthcare of Atlanta (CHOA) pediatric ICU database. The imported data contained information collected in 5,739 ICU stays from 4,975 patients aged from birth to 21 years old in the year of 2013. The data can be categorized into four major categories, including visit information, procedures, laboratory testing, and microbiology testing. Other than visit information, all data was collected with timestamps, which enabled the mining of temporal association rules. Examples and number of records in each category are tabularized in Table 2.

3.2 Effect of rule components

3.2.1 Lab testing vs. two-hour ICU mortality—The prediction of the ICU mortality has been widely studied using conventional scales such as the admission-based APACHE-II [3] or the daily-based Sequential Organ Failure Assessment (SOFA) [7]. However, risk prediction models using laboratory (lab) testing are relatively rare even though they also frequently occur in the ICU setting [34]. Developing risk prediction models by including lab testing allows us to utilize hundreds of clinical attributes instead of a fixed number of items as in conventional scales (e.g., 12 routine physiologic measures and six basic scores in SOFA). Large volume of clinical attributes are fundamental for personalized risk prediction since they can comprehensively cover individual characteristics. Therefore, in this case study, we employed icuARM-II to demonstrate its ability of short-term (i.e., 2-hr) ICU mortality prediction based on personalized lab testing results.

The lab testing dataset in CHOA ICU database consists of more than three million records from more than one thousand tests. We selected the top 12 most counted tests and converted the numerical values into either abnormal or normal levels based on the suggested ranges. The total number of records in each test and those in each level are listed in Table 3. To predict the 2-hr ICU mortality based on personalized lab testing results, the rule was in the form of $A \Rightarrow C$: $\{E_A \mid T_A\} \Rightarrow \{Death \mid <2\text{-}hr\}$. The antecedent episode A represents a set of abnormal lab testing results E_A that have been observed in the last time period with a length of T_A to predict the *Death* event in the following two hours. The prediction possibility was then determined by the maximum confidence value of the generated rules in the decision list. Since the abnormality of a lab test can occur multiple times within a observation period, we simplified E_A with a chain of ExN in which E was a lab testing item and N indicated its repeat. For instance, if a patient has had abnormal glucose level twice and abnormal creatinine level once in the past day, the rule was represented as $\{GLU$x$2, CRE$x$1 \mid <1\ day\} \Rightarrow \{Death \mid <2\text{-}hr\}$. Throughout the mining process in this case study, we apply $Supp_{min}$ = 0.5% to detect rare episodes and $Conf_{min}$ = 5% to capture the death possibility as low as 5%.

3.2.2 Quality assessment of lab tests—Before mining rules for patients, we can assess and rank the reliability of individual lab tests in the prediction of 2-hr ICU mortality. Since a lab test can occur multiple times within an observation period, we first generated reliabilities of rules that share the same lab test but with repeats from one to a number *N*. Then we calculated the average of reliability values from these *N* rules for the overall quality of this lab test. For example, we assessed the quality of the glucose test in one day by calculating the average reliability value of rules $\{GLUxN \mid <1\ day\} \Rightarrow \{Death \mid <2\text{-}hr\}$ where *N* ranged from one to a certain repeat. Then we could change the glucose test to a creatinine test with the same range of repeat and compare its result with the glucose test. In addition, for each decision list generated given one lab testing item and a repeat, we compared the performance from our voting-based classification approach with it from the single rule-based approach in the CBA approach. We applied this quality assessment on the 12 selected lab tests with ranges from one to 10.

The results of the quality assessment are shown in Figure 5. Generally, the performance of our classification approach outperforms CBA by comparing the reliability to the CBA's AUC values. In addition, a prediction model should have at least 80% of AUC to be considered reliable according to medical standard [35]. The quality assessment results generated by our association-based classifier show that only three lab tests (i.e., arterial pH (*APH*), PO_2 (*PO2*), and glucose (*GLU*)) were qualified for the 2-hr ICU mortality prediction if they have been observed individually within one day. Meanwhile, among all other nine lab tests with reliabilities <80%, total CO_2 (*CO2*), arterial base excess (*ABE*), and creatinine (*CRE*) are the worst three lab tests for the prediction of 2-hr ICU mortality.

3.2.3 Interactions of lab tests—The assessment task allowed us to evaluate the quality of individual lab tests. However, in the real ICU setting, it is very common that patients receive different lab tests over the course of one day. Therefore, it is worth investigating the interaction among lab tests even though they may perform worse individually. We assumed that the possibility (i.e., confidence) of ICU mortality could increase when more abnormal lab test results had been observed. Additionally, the prediction reliability could also be improved if more lab tests (i.e., more information) were available.

To investigate the effect of lab testing interaction, we randomly selected four lab tests (i.e., *CA*, *SOD*, *CO2*, and *ABE*). The investigation started from the *CA* test and added other tests one-by-one. Each test was abnormal twice. Thus the four rules were compared:

$$\{CAx2 \mid <1\text{-}day\} \Rightarrow \{Death \mid <2\text{-}hr\}$$

$$\{SODx2 + CAx2 \mid <1\text{-}day\} \Rightarrow \{Death \mid <2\text{-}hr\}$$

$$\{CO2x2 + SODx2 + CAx2 \mid <1\text{-}day\} \{Death \mid <2\text{-}hr\}$$

$$\{ABEx2 + SO2x2 + SODx2 + CAx2 \mid <1\text{-}day\} \Rightarrow \{Death \mid <2\text{-}hr\}$$

According to the results shown in Figure 6, if a patient was observed with abnormal ionized calcium twice (i.e., *CAx2*) in the past one day, the possibility of death in the following two hours was 5.0% with the prediction reliability (i.e., AUC) of 73.9%. Both the confidence and reliability increased monotonically when more abnormal tests were observed. If a patient

had twice as many abnormalities in all *CA*, *SOD*, *CO2*, and *ABE* in one day, the possibility of death in the following two hours increased to 14.9% with reliability of 90.2%. Therefore, our results verified our assumptions that the possibility (i.e., confidence) of 2-hr mortality was increased with interactions among more abnormal lab testing results, which also improved the reliability of the prediction.

3.2.4 Additional lab tests—So far a clinician can predict a patient's short-term ICU mortality using the observed abnormal lab testing results. However, given this set of observed abnormalities, the clinician usually wants to determine what other abnormalities can be used to confirm the prediction and to avoid the increase of the possibility of death by giving proper treatments. For example, if a patient is observed with abnormal ionized calcium twice (*CA*x*2*) in the past 24 hours, according to the results of Section 3.2.3, we know that the possibility of death in the following 2-hr is 5.0% with the prediction reliability of 73.9%. We would like to decide which test is additionally necessary to better predict the short-term mortality for this patient and what the proper treatments can be offered to decrease the possibility. Assuming the remaining 11 lab tests other than *CA* are considered, we can evaluate the confidence and reliability values from 11 rules and each of which has *CAx2* plus once of an abnormal lab test. According to Table 4, adding abnormal creatinine once (*CREx1*) to *CAx2* can increase the confidence the most from 5.0% to 8.5% with reliability improved from 73.9% to 88.9%. Therefore, the clinician should order a lab test for creatinine to confirm and to provide proper treatments for its abnormality to avoid the short-term mortality. Offering such information can prevent unnecessary lab tests that are ordered by default following clinical guidelines, instead of being driven by patient characteristics, to prevent high healthcare costs and human-based errors [36, 37].

3.2.5 Length of the observation period—The same set of abnormal lab tests observed in different lengths of observation periods may also be associated with different risks of mortality and/or the corresponding prediction reliabilities. To investigate the effect of observation length, we continued using all of the four lab tests (i.e., *CA*, *SOD*, *CO2*, and *ABE*) that were selected in the Section 3.2.3. Assuming each lab test has been observed abnormal twice, we consider the following five rules with five different observation lengths ranging from half day to four days:

$$\{\text{ABEx2} + \text{SO2x2} + \text{SODx2} + \text{CAx2} \mid <\text{½-day}\} \Rightarrow \{\text{Death} \mid <\text{2-hr}\}$$

$$\{\text{ABEx2} + \text{SO2x2} + \text{SODx2} + \text{CAx2} \mid <\text{1-day}\} \Rightarrow \{\text{Death} \mid <\text{2-hr}\}$$

$$\{\text{ABEx2} + \text{SO2x2} + \text{SODx2} + \text{CAx2} \mid <\text{2-day}\} \Rightarrow \{\text{Death} \mid <\text{2-hr}\}$$

$$\{\text{ABEx2} + \text{SO2x2} + \text{SODx2} + \text{CAx2} \mid <\text{3-day}\} \Rightarrow \{\text{Death} \mid <\text{2-hr}\}$$

$$\{\text{ABEx2} + \text{SO2x2} + \text{SODx2} + \text{CAx2} \mid <\text{4-day}\} \Rightarrow \{\text{Death} \mid <\text{2-hr}\}$$

The results in Figure 7 suggest that the same set of abnormal conditions happened more recently may be associated with a higher possibility of short-term mortality. If a patient has been observed abnormal in *CA*, *SOD*, *CO2*, and *ABE* that each has two occurrences in the past half day, the patient's possibility of death in the following two hours is 16.2%, and the reliability of this prediction is 90.8%. If another patient had been observed with the same set of abnormality but in the past four days, the possibility of 2-hr mortality is 9.5% with

reliability of 83.4%. Therefore, our results show that abnormal conditions observed more recently may be associated with a higher chance of short-term mortality. In addition, a shorter observation period slightly improves the prediction reliability.

3.3 Discussion

Based on the results in Section 3.2, we can conclude that factors of the increase of abnormality repeat, interaction of different abnormalities, and decrease of the observation length all have positive impacts on the possibility of short-term ICU mortality and the corresponding prediction reliability.

We should note that all of the rules generated in Section 3.2 are examples to understand the influences of changes in the three aforementioned factors. Clinicians do not need to adopt these rules as generalized or predefined evidence for the decision making. Instead, based on a patient's instant conditions, clinicians can construct a personalized antecedent episode and a target consequent episode with flexible time windows. In addition, in this case study we only used 12 lab-testing items to demonstrate the rule mining framework. The framework is actually capable of more than 5,000 lab tests, 1,000 microbiology tests, 250 clinical procedures, and more than 40 basic information regarding ICU visits. To enable the usability of icuARM-II, we have developed a friendly user interface to enable the real-world clinical decision making, which is introduced in the following section.

4. User Interface

The icuARM-II features a user interface that allows real-time ICU mortality prediction. The interface was implemented in MATLAB (MathWorks, Natick, MA). As shown on the top of Figure 8, the operation starts from constructing new events by selecting a category, choosing a procedure and event (e.g., lab measures) under the category, assigning the number of repeat of the event, specifying a length of the observation window, and providing the target length of prediction period. The constructed events represent the current antecedent episode, which are shown in the Current Episode panel of Figure 8. Users can also manage the current antecedent episode by adding or removing events. The Run button triggers the prediction process. Based on the given patient episode and target mortality prediction period, the interface displays the final possibility with the performance information, including the ROC curve and the reliability value.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose an ICU decision support system, called icuARM-II, to provide flexible clinical risk prediction based on personalized temporal conditions. This study is an extension of our previous work on ICU data mining, which was restricted to non-temporal clinical data, using a newly developed pediatric ICU database from Children's Healthcare of Atlanta (CHOA). We first introduced a scanning strategy to count temporal association rules. We then applied classification-based rule generation to produce a decision list. Given a decision list, we proposed a classification approach, using the principle of leave-one-out cross validation, to calculate a new reliability metric. The proposed framework was tested using the lab testing dataset for the prediction of short-term (i.e., 2-hr) ICU mortality. Our

results not only outperform conventional Classification Based on Association (CBA), but also suggest important usability regarding lab testing, including (1) quality assessment of individual tests, (2) evaluation of interactions among tests, (3) suggestion of additional tests, and (4) investigating the change of observation length. Featuring with interactive user interface, icuARM-II has demonstrated a new solution for real-time and reliable risk prediction using personalized clinical data.

While the current results of this study are promising, we also plan to improve icuARM-II in the following five directions. Firstly, we plan to compare the prediction performance of our rule mining framework to other classification-based temporal and sequential mining approaches that were designed to discover hidden relations between sequences and sub-sequences of events. The target approaches include hidden Markov model (HMM) [38] and one-nearest neighbor classifier with semi-supervised time series classification [39]. We will also consider other metrics, such as classification-based calibration [40, 41], to evaluate and ensure the usefulness of the model. Secondly, other than lab testing results, we would incorporate variables from other categories in the CHOA ICU database, such as more comprehensive ICU stay information, microbiology testing, and procedures. By including this breadth of data, we may be able to better emulate many factors that impact each decision made by clinicians. Thirdly, the current icuARM-II only leverages data collected from 5,739 ICU visits during a one-year (i.e., 2013) period for the purpose of methodology development and evaluation. We will import more ICU data that can expand to 10 years back from now. This future work will improve the generalizability of our findings. Fourthly, icuARM-II still has a bottleneck with a long latency from giving input to obtaining results. Therefore, we are investigating solutions to improve icuARM-II's entire efficiency, from rule generation to reliability assessment, which will be important after we import more ICU clinical data. Finally, we will include the prediction of other clinical risks (e.g., prolonged ICU stays and ventilator days) with better-designed user interface to maximize the potential of clinical decision support using icuARM-II.

Acknowledgments

We thank Matthew Miller, Michael Thompson, Sherry Farrugia, and Tod Davis for helping the development of the CHOA ICU database used in this study. The authors are also grateful to Dr. Sonal Kothari, Chanchala Kaddi, Theruni Pethiyagoda, and Po-Yen Wu for their valuable comments and suggestions. This research has been supported by grants from NIH (U54CA119338, 1RC2CA148265, and R01CA163256), Georgia Cancer Coalition Award to Prof. MD Wang, Hewlett Packard, and Microsoft Research.

References

1. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review. 1956; 63(2):81. [PubMed: 13310704]
2. Rosenberg AL. Recent innovations in intensive care unit risk-prediction models. Current opinion in critical care. 2002; 8(4):321–330. [PubMed: 12386493]
3. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Critical care medicine. 1985; 13(10):818–829. [PubMed: 3928249]
4. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. Jama. 1993; 270(20): 2478–2486. [PubMed: 8230626]

5. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. Jama. 1993; 270(24):2957–2963. [PubMed: 8254858]
6. Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Critical care medicine. 1995; 23(10):1638–1652. [PubMed: 7587228]
7. Vincent JL, Ferreira F, Moreno R. Scoring systems for assessing organ dysfunction and survival. Critical care clinics. 2000; 16(2):353–366. [PubMed: 10768086]
8. Le Gall JR, Klar J, Lemeshow S, Saulnier F, Alberti C, Artigas A, Teres D. The Logistic Organ Dysfunction system: a new way to assess organ dysfunction in the intensive care unit. Jama. 1996; 276(10):802–810. [PubMed: 8769590]
9. Fialho AS, Cismondi F, Vieira SM, Reti SR, Sousa JM, Finkelstein SN. Data mining using clinical physiology at discharge to predict ICU readmissions. Expert Systems with Applications. 2012; 39(18):13158–13165.
10. Clark PA, Lettieri CJ. Clinical model for predicting prolonged mechanical ventilation. Journal of critical care. 2013; 285:880.e881–880.e887. [PubMed: 23683556]
11. Elstein AS. Heuristics and biases: selected errors in clinical reasoning. Academic Medicine. 1999; 74(7):791–794. [PubMed: 10429587]
12. Zygun DA, Laupland KB, Fick GH, Sandham JD, Doig CJ. Neuroanesthesia and Intensive Care Limited ability of SOFA and MOD scores to discriminate outcome: a prospective evaluation in 1,436 patients. Canadian Journal of Anesthesia. 2005; 52(3):302–308. [PubMed: 15753504]
13. Khwannimit B. A comparison of three organ dysfunction scores: MODS, SOFA and LOD for predicting ICU mortality in critically ill patients. Medical Association of Thailand, Journal of. 2007; 90(6):1074.
14. Holtfreter B, Bandt C, Kuhn SO, Grunwald U, Lehmann C, Schütt C, Gründling M. Serum osmolality and outcome in intensive care unit patients. Acta anaesthesiologica scandinavica. 2006; 50(8):970–977. [PubMed: 16923092]
15. Timsit JF, Fosse JP, Troché G, de Lassence A, Alberti C, Garrouste-Orgeas M, Bornstain C, Adrie C, Cheval C, Chevret S. Calibration and discrimination by daily Logistic Organ Dysfunction scoring comparatively with daily Sequential Organ Failure Assessment scoring for predicting hospital mortality in critically ill patients*. Critical care medicine. 2002; 30(9):2003–2013. [PubMed: 12352033]
16. Cheng CW, Chanani N, Venugopalan J, Maher K, Wang D. icuARM–An ICU Clinical Decision Support System Using Association Rule Mining. Translational Engineering in Health and Medicine (JTEHM), IEEE Journal of. 2013; 1
17. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems. 2012; 36(4): 2431–2448. [PubMed: 21537851]
18. Agrawal, R.; Imieliński, T.; Swami, A. Proceedings of the ACM SIGMOD Record. ACM; 1993. Mining association rules between sets of items in large databases; p. 207-216.
19. Agrawal, R.; Srikant, R. Proceedings of the Proc 20th int conf very large data bases, VLDB. ACM; 1994. Fast algorithms for mining association rules; p. 487-499.
20. Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics. 2003; 19(1):79–86. [PubMed: 12499296]
21. Leung KS, Wong KC, Chan TM, Wong MH, Lee KH, Lau CK, Tsui SK. Discovering protein–DNA binding sequence patterns using association rule mining. Nucleic acids research. 2010; 38(19):6324–6337. [PubMed: 20529874]
22. Konias, S.; Giaglis, G.; Gogou, G.; Bamidis, P.; Maglaveras, N. Proceedings of the Computers in Cardiology. IEEE; 2003. Uncertainty rule generation on a home care database of heart failure patients; p. 765-768.
23. Ordonez C, Omiecinski E, de Braal L, Santana CA, Ezquerra NF, Taboada JA, Cooke CD, Krawczynska E, Garcia EV. Mining Constrained Association Rules to Predict Heart Disease. Proceedings of the nternational Conference on Data Mining (ICDM). 2001:433–440.

24. Shan, Y.; Jeacocke, D.; Murray, DW.; Sutinen, A. Proceedings of the Proceedings of the 7th Australasian Data Mining Conference-Volume 87. Australian Computer Society, Inc; 2008. Mining medical specialist billing patterns for health service management; p. 105-110.
25. Bellazzi R, Larizza C, Magni P, Bellazzi R. Temporal data mining for the quality assessment of hemodialysis services. Artificial Intelligence in Medicine. 2005; 34(1):25–39. [PubMed: 15885564]
26. Chaves R, Górriz J, Ramírez J, Illán I, Salas-Gonzalez D, Gómez-Río M. Efficient mining of association rules for the early diagnosis of Alzheimer's disease. Physics in medicine and biology. 2011; 56(18):6047. [PubMed: 21873769]
27. Cheng, CW.; Burns, TG.; Wang, MD. Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2013. Mining Association Rules for Neurobehavioral and Motor Disorders in Children Diagnosed with Cerebral Palsy; p. 258-263.
28. Cheng, CW.; Martin, GS.; Wu, PY.; Wang, MD. Proceedings of the The International Conference on Health Informatics. Springer; 2014. PHARM-Association Rule Mining for Predictive Health; p. 114-117.
29. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Critical care medicine. 2011; 39(5):952. [PubMed: 21283005]
30. Verma, K.; Vyas, OP.; Vyas, R. Temporal approach to association rule mining using t-tree and p-tree. Springer; City: 2005.
31. Ale, JM.; Rossi, GH. An approach to discovering temporal association rules. ACM; City: 2000.
32. Provost FJ, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. Proceedings of the ICML. 1998:445–453.
33. Ma BLWHY. Integrating classification and association rule mining. Proceedings of the Proceedings of the 4th. 1998
34. Roberts DE, Bell DD, Ostryzniuk T, Dobson K, Oppenheimer L, Martens D, Honcharik N, Cramp H, Loewen E, Bodnar S. Eliminating needless testing in intensive care-an information-based team management approach. Critical care medicine. 1993; 21(10):1452–1458. [PubMed: 8403952]
35. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients*. Critical care medicine. 2006; 34(5):1297–1310. [PubMed: 16540951]
36. Mehari S, Havill J. Written guidelines for laboratory testing in intensive care-still effective after 3 years. Critical Care and Resuscitation. 2001; 3(3):158. [PubMed: 16573496]
37. Garland A, Shaman Z, Baron J, Connors AF Jr. Physician-attributable differences in intensive care unit costs: a single-center study. American journal of respiratory and critical care medicine. 2006; 174(11):1206–1210. [PubMed: 16973977]
38. Zhong S. Semi-supervised sequence classification with hmms. International Journal of Pattern Recognition and Artificial Intelligence. 2005; 19(02):165–182.
39. Wei, L.; Keogh, E. Proceedings of the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. Semi-supervised time series classification; p. 748-753.
40. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. ACM; City: 2002.
41. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. Journal of the American Medical Informatics Association. 2012; 19(2): 263–274. [PubMed: 21984587]

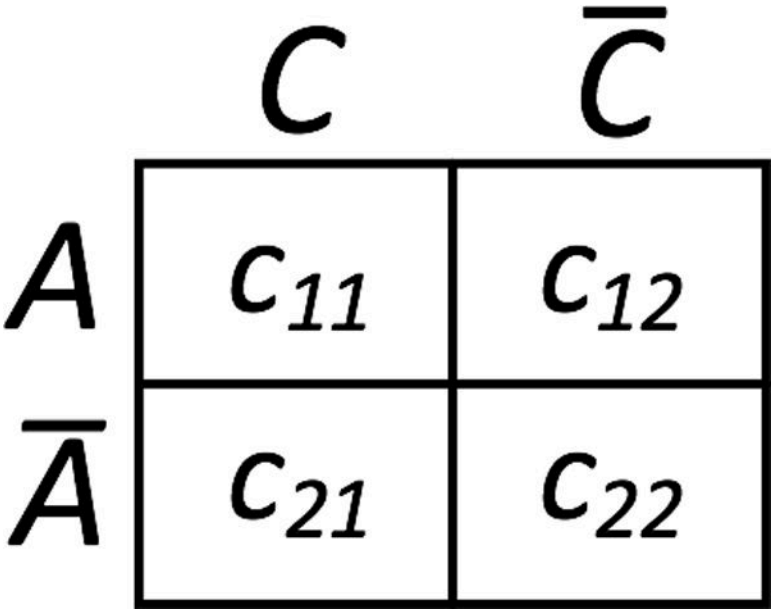


Figure 1.
Contingency table of a temporal association rule $A \Rightarrow C$.

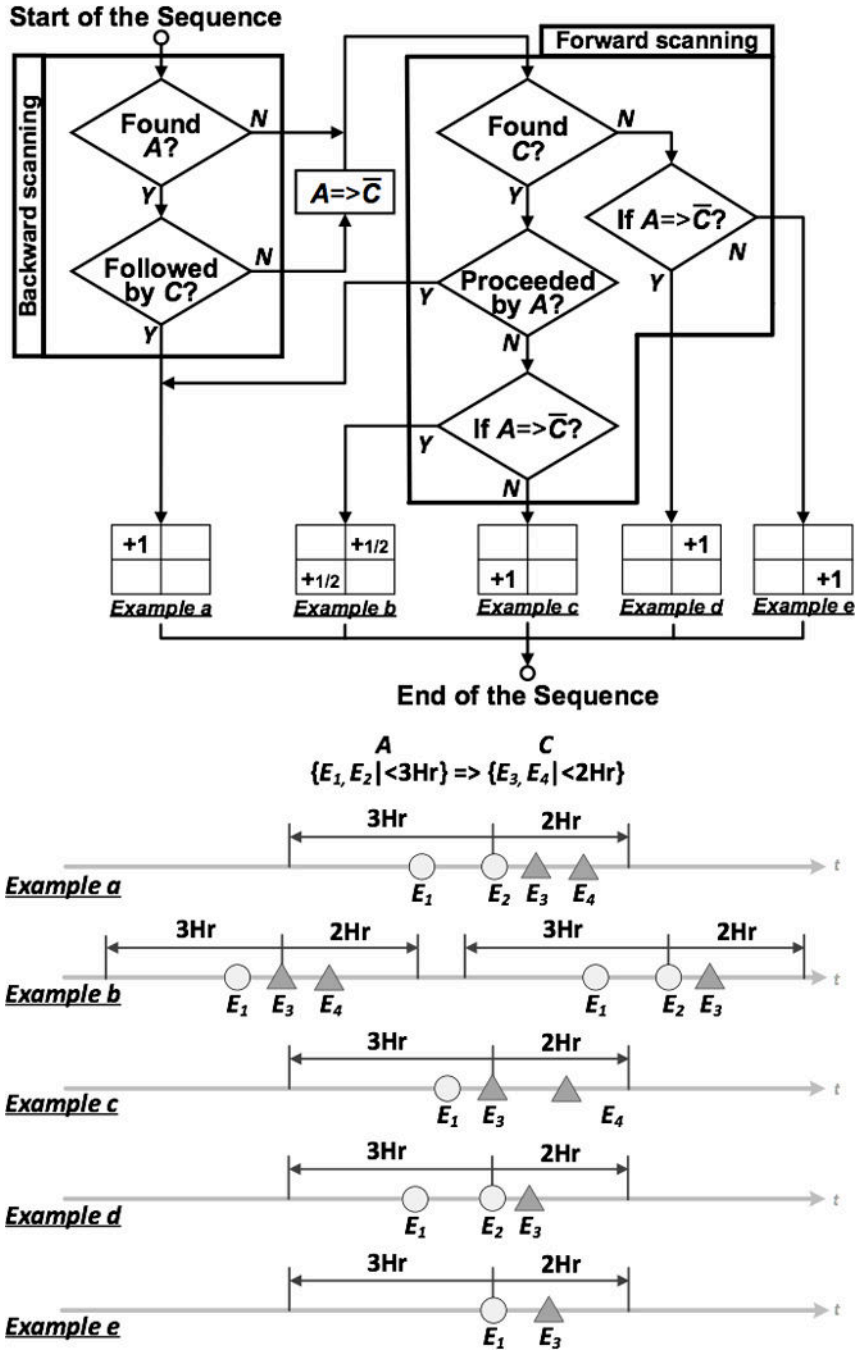


Figure 2.
Flow of backward and forward scanning (top) with examples of five different situations (bottom).

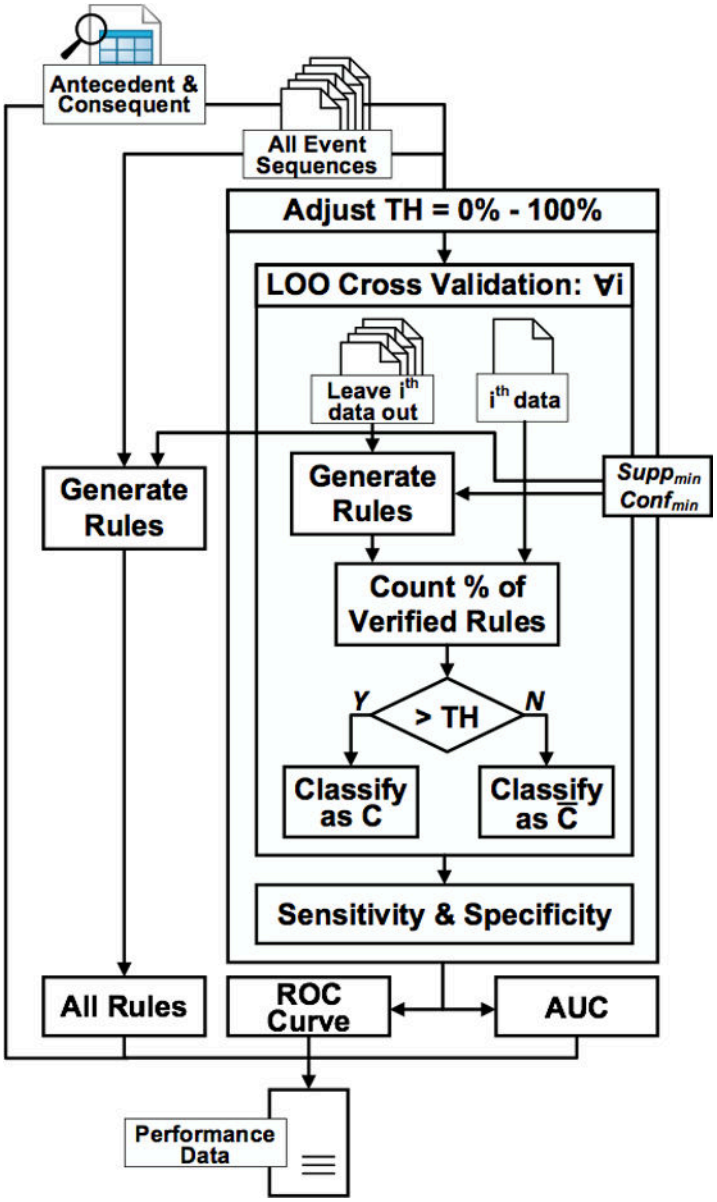


Figure 3.
Steps of the generation of performance data.

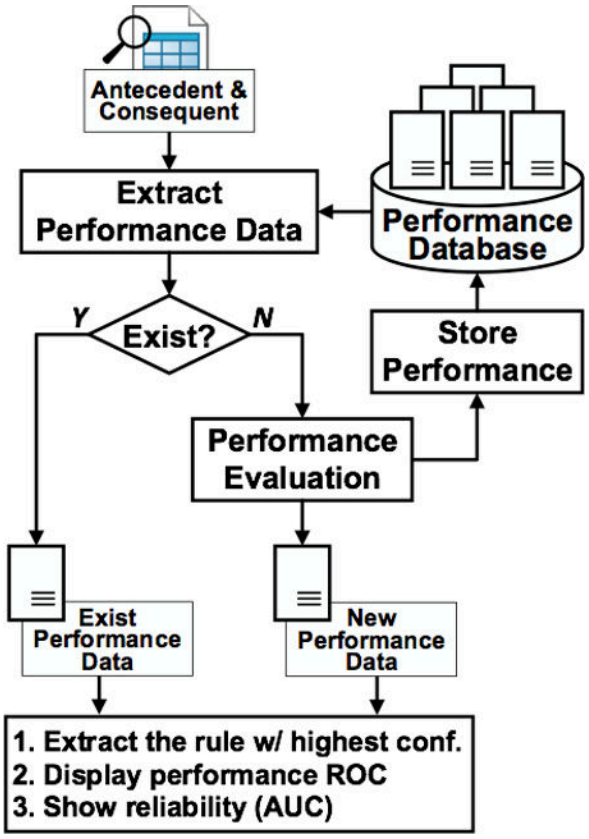


Figure 4.
Flow of the creation, storage, and retrieval of rule performance data.

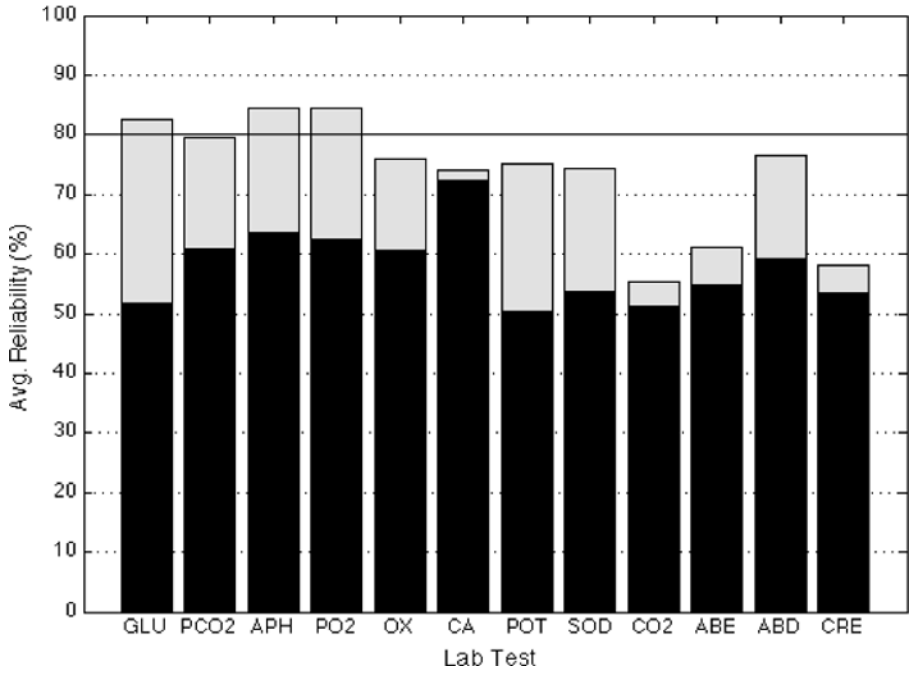


Figure 5.
Average reliabilities (light gray) and CBA performance (black) of 12 selected lab tests.

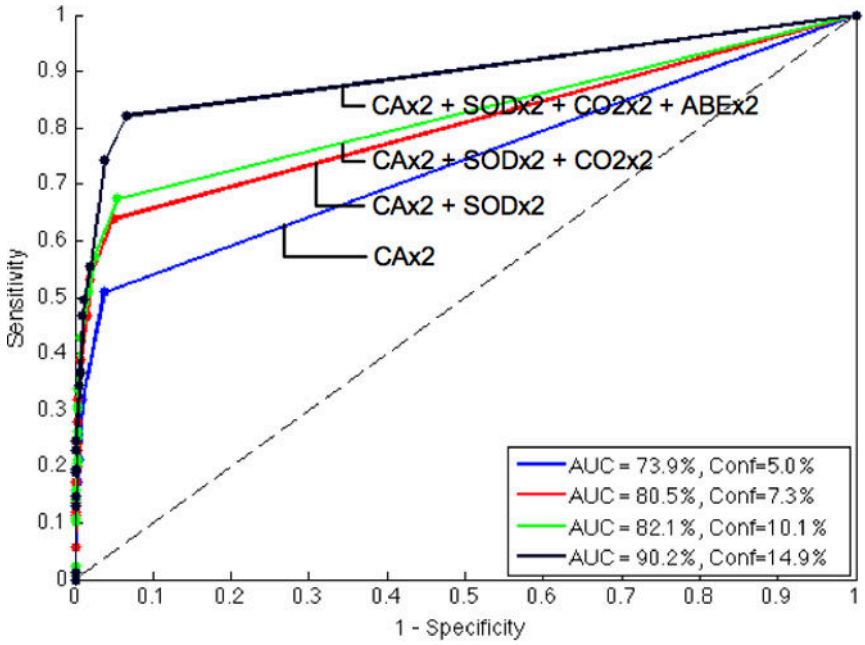


Figure 6.
Effects of interactions among four lab tests.

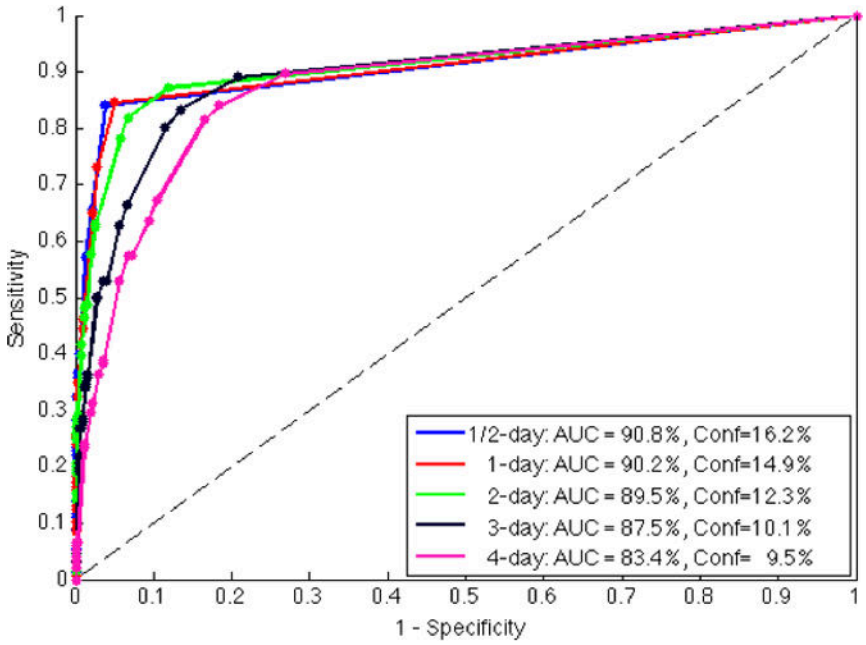


Figure 7.
Effects of different observation lengths.

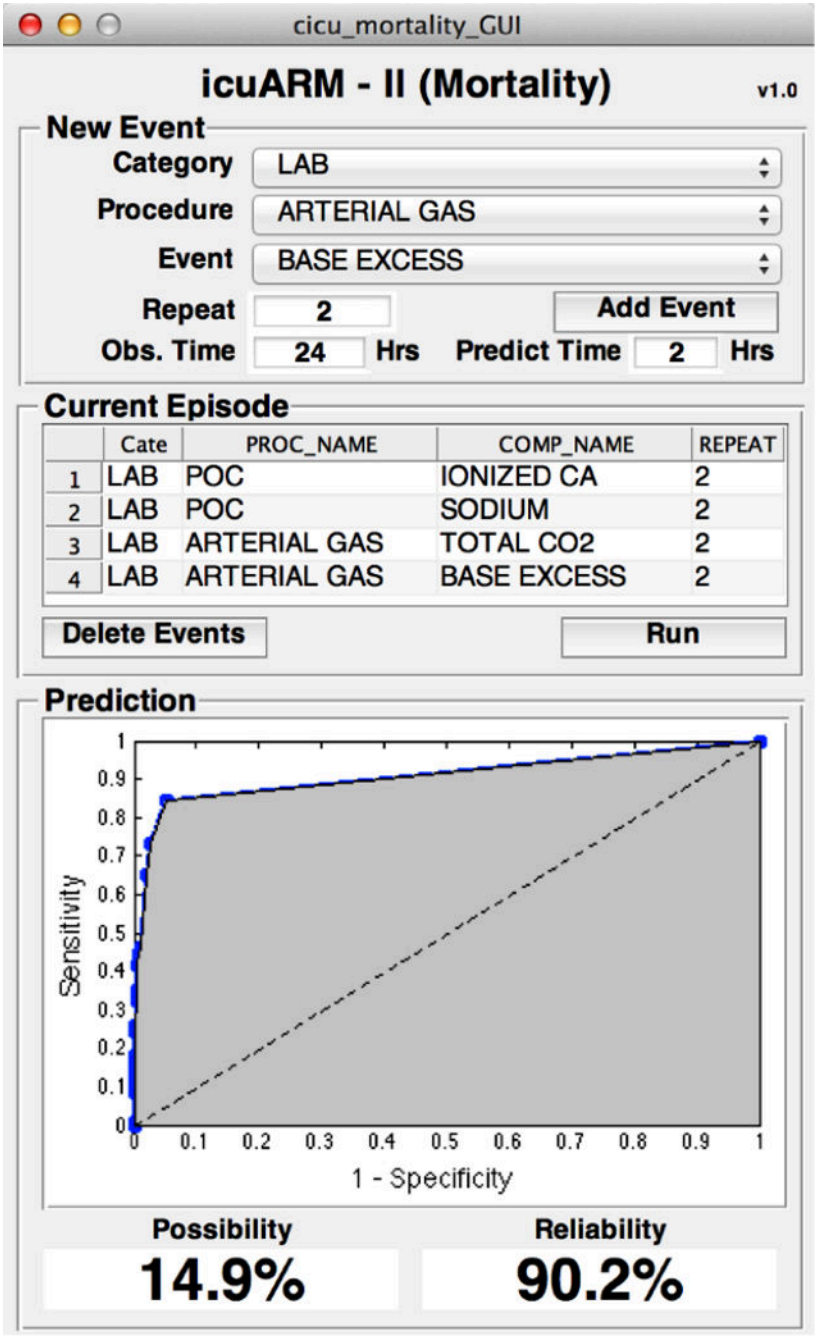


Figure 8.

The user interface of icuARM-II for ICU mortality prediction. It demonstrates the setting and the result of a rule: {*ABEx2* + *SO2x2* + *SODx2* + *CAx2* /<*1-day*}⇒ {*Death*/ <*2-hr*}

Table 1

Use cases and examples with different settings of T_A and T_C in temporal association rules

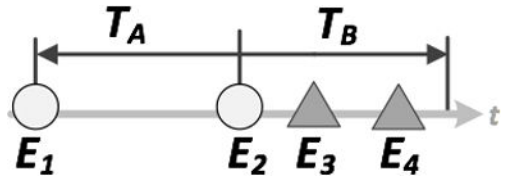
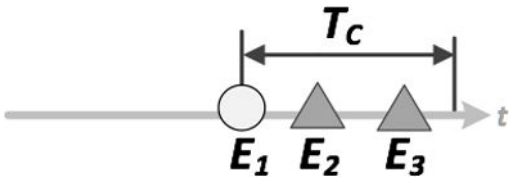
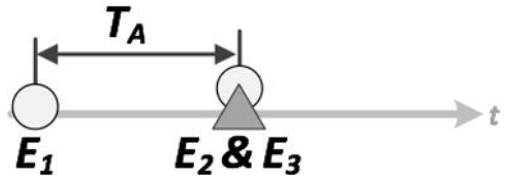
Setting	Illustration	Prediction Example
$T_A \neq 0$, $T_C \neq 0$		After finishing two treatments E_1 and E_2 in a period of T_A, what is the possibility of the development *of high heart rate* (E_3) and *low arterial pH* (E_4) in the following T_C?
$T_A = 0$, $T_C \neq 0$		After taking a drug E_1, what is the possibility of the development of *high blood pressure* (E_2) and *high creatinine level* (E_3) in the following T_C?
$T_A \neq 0$, $T_C = 0$		Upon finishing of two treatments E_1 and E_2 in a period of T_A, what's is the possibility of development of *low pulse oximetry* (E_3) at the end of T_A?
$T_A = 0$, $T_C = 0$		What is the possibility that *low white blood cell counts* (E_1) coexists with *low urine output* (E_2)?

Table 2

Categories and examples of CHOA ICU Database

Category	Measure Examples	# of records
Visit Info.	Demographics, admission/discharge time, birth weight/length, discharge destination, APGAR 1, 5, 10 minutes score, ventilator days, financial class, PICU, NICU, CICU flags	5,738
Procedures *	Oxygen supply, aerosol treatment, oxygen per shift, PH probe, pulse ox assessment, gastric pressure, suction, reactive protein, tobramycin peak	416,520
Lab Testing *	Glucose, arterial PCO_2/pH/PO_2, oxygen saturation, calcium ionized, HCO_3, creatinine, platelet count, potassium, prolactin, salicylates,	3,348,924
Microbiology Testing *	Culture, specimen description, specimen source	87,843

* Temporal data

Table 3

Top 12 counted lab testing items in CHOA ICU database

Lab Item	ID	# of Records		
		Normal	Abnormal	Total
Glucose	*GLU*	19,793	38,828	58,621
Arterial PCO_2	*PCO2*	26,898	22,348	49,246
Arterial pH	*APH*	21,772	27,474	49,246
Arterial PO_2	*PO2*	5,733	43,513	49,246
Ox Saturation	*OX*	22,035	27,211	49,246
Ionized Ca	*CA*	28,690	14,484	43,174
Potassium	*POT*	24,232	16,693	40,925
Sodium	*SOD*	24,870	14,854	39,724
Total CO_2	*CO2*	42,898	6,348	49,246
Art. Base Excess	*ABE*	11,198	14,502	25,700
Art. Base Deficit	*ABD*	8,533	15,062	23,595
Creatinine	*CRE*	18,607	3,263	21,870
			Total	499,839

Table 4

Effects of additional lab tests

1st Test	2nd Test	Confidence	Reliability
CAx2	∅	5.0%	73.9%
	GLUxl	5.2%	85.2%
	PCO2xl	6.1%	88.1%
	APHxl	6.2%	90.0%
	PO2xl	5.7%	90.5%
	OXxl	8.1%	85.6%
	CO2xl	8.1%	81.4%
	POTxl	5.6%	80.4%
	SODxl	5.6%	76.2%
	ABExl	5.0%	74.0%
	ABDxl	6.9%	85.3%
	CRExl	**8.5%**	88.9%