Published in final edited form as:

ACM BCB. 2015 September ; 2015: 386–392. doi:10.1145/2808719.2808759.

Improving Personalized Clinical Risk Prediction Based on Causality-Based Association Rules

Chih-wen Cheng and May D. Wang

Abstract

Developing clinical risk prediction models is one of the main tasks of healthcare data mining. Advanced data collection techniques in current Big Data era have created an emerging and urgent need for scalable, computer-based data mining methods. These methods can turn data into useful, personalized decision support knowledge in a flexible, cost-effective, and productive way. In our previous study, we developed a tool, called icuARM- II, that can generate personalized clinical risk prediction evidence using a temporal rule mining framework. However, the generation of final risk prediction possibility with icuARM-II still relied on human interpretation, which was subjective and, most of time, biased. In this study, we propose a new mechanism to improve icuARM-II's rule selection by including the concept of causal analysis. The generated risk prediction is quantitatively assessed using calibration statistics. To evaluate the performance of the new rule selection mechanism, we conducted a case study to predict short-term intensive care unit mortality based on personalized lab testing abnormalities. Our results demonstrated a better-calibrated ICU risk prediction using the new causality-base rule selection solution by comparing with conventional confidence-only rule selection methods.

1. INTRODUCTION

Advanced information technologies promise the massive influx of individual-centered clinical data. These rich sources of data provide the potential for an increased understanding of disease mechanisms and patient-centered risk prediction to reduce unnecessary cost without sacrificing the quality of care. However, the size, complexity, and biases of the data pose new challenges, which makes it difficult to transform the data to reliable risk prediction knowledge by using conventional statistical analyses. To obtain patient-specific clinical risk prediction at the point-of- care, it is necessary to introduce advanced, computer-based data mining methods and tools for providers to interpret the information in the context of that patient. Such personalized medicine represents the customization of care for each patient [1]. Even though studies have suggested the potential of such models to offer more reliable patient-specific predictions [2], studies with real-time model development and demonstration remain very rare.

Besides, current information technologies can collect longitudinal clinical data throughout a medical encounter, e.g., data collected by bedside monitors. It is becoming more important

*Corresponding Author, Contact information for the 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.

to consider the characteristics, which constantly change with time, and develop temporal knowledge-mining frameworks to uncover the essence of clinical episode evolution. Furthermore, the area under the receiver operation characteristic curve (AUC), or the c -statistic, is the most popular tool for assessing the quality of clinical prediction models. However, using AUC statistic in the prediction of future events remains imperfect [3, 4].

Temporal association rule mining has been studied in many researches [5, 6]. However, models in these studies tends to generate all possible rules from the dataset, instead of generating only rules that can precisely describe an individual's clinical conditions. In our previous study, we developed a tool, called icuARM-II, with a mining framework that can generate temporal association rules for personalized risk prediction in the intensive care unit (ICU) [7]. However, icuARM-II's mining framework still tends to generate 'a list' of rules with different prediction possibilities that are all applicable to a patient's clinical conditions. Our users still need to manually determine the final, single, prediction possibility from the rule list, which is often time-consuming, clueless, and subject. Thus, the question remained unsolved is how to automatically and objectively generate a final prediction possibility from a rule list.

In this study, we aimed to improve icuARM-II's rule selection mechanism by combining the concept of causal analysis. In the first part of the following sections, we first introduce the basic concept of association rule mining, causal analysis, and our newly proposed rule selection mechanism. Then, in Section 3, we adopt calibration assessment to quantitatively measure the performance of the new causality-based rule selection. Then we conduct a case study, in Section 4, to demonstrate the improvement of rule selection in the prediction of short-term ICU mortality using temporal lab testing abnormalities. Finally, we provide the conclusion and future directions in Section 5.

2. CAUSAL-BASED RULE SELECTION

2.1 Association Rule Mining

Association rule mining (ARM) is one of the data mining methods to discover meaningful relations between variables in databases. Agrawal *et al.*, first introduced the concept of ARM for extracting regularities between sales items in large-scale warehouse databases [8]. An association rule is in the form of $X \Rightarrow Y$ to represent the relationships between the antecedent (X) and the consequent (Y) [9]. In clinical data mining, if the most basic data tuple is a patient's clinical report collected in one visit, the rule $X \Rightarrow Y$ implies that if conditions in X occur during the visit, another set of conditions in Y are also likely to occur during the same visit. For example, a rule $\{SerumCholesterol > TH_{SC} \& SerumGlucose > TH_{SG}\} \Rightarrow \{RespiratoryFailure = High\}$ implies that "in one clinical stay, if a patient's serum cholesterol and serum glucose have been elevated above two thresholds, TH_{SC} and TH_{SG} , the patient's risk of respiratory failure may also be increased."

One of the metrics to assess an association rule is called support, which is calculated as:

$$\text{support}(X \Rightarrow Y) = \frac{\text{count}(X \cup Y)}{\text{Total \# of tuples in database}} \quad (1)$$

where $\text{count}(X \cup Y)$ returns the number of transactions that contain both X and Y . A high support of an association rule means that the rule is applicable to a large portion of the data set. The other metric, confidence, indicates how much Y appears in those tuples that already contain X :

$$\text{confidence}(X \Rightarrow Y) = \frac{\text{count}(X \cup Y)}{\text{count}(X)} \quad (2)$$

For example, in clinical outcome prediction, if an association rule has confidence of 85%, this value implies that if a patient has developed a set of clinical events X , he or she may have 85% of chance to develop another set of outcomes Y . In other words, confidence value reveals the level of the association between X and Y . Afterwards, the mining process requires users to specify a minimum support (Supp_{min}) and a minimum confidence (Conf_{min}) to eliminate infrequent and unconfident rules, respectively. Details regarding the generation of frequent itemsets and confident rules can be referred in [10].

ARM has been widely adopted in healthcare settings. For example, in the prediction of heart disease, Konias *et al.*, proposed an uncertainty rule generator (URG) to discover rules for home-care monitoring from congestive heart failure patients [11]. Auditing medical abusive and fraudulent behavior is another important application of ARM. Shan *et al.*, used ARM to examine medical billing patterns to detect suspicious claims and potentially fraudulent individuals [12]. Similarly, Bellazzi *et al.*, introduced a temporal ARM framework in an auditing system to facilitate the understanding of patients' behavior and improve the quality of hemodialysis services [13]. ARM also demonstrates successfulness in neurology research. For example, authors in [14] proposed an ARM mining flow to discover image-based association rules in functional single-photon emission computed tomography (SPECT) image databases to improve early diagnosis of Alzheimer's disease. Studies also have verified that ARM can find hidden diagnosis rules of developmentally-delayed children, so as to enable healthcare professionals in early intervention of delayed psychological developments [15].

2.2 Causal Analysis

One of the disadvantages of ARM is that the mined association rules can be spurious and do not reflect true causality between the antecedent and consequent. Thus association rule mining cannot be called 'causal' analysis because the rule cannot imply the relation between an antecedent (the cause) and a consequent (the effect), where the effect is understood as a physical consequence of the cause. This means that many of mined rules may not have practical meaning if not being verified by human knowledge. While the development of reliable mining process for finding causality patterns in clinical data, the determination of real causes given a target outcome has become prominent.

Causal relationships imply how the outcome would be affected when the cause is changed. The gold standard for conventional causal analysis is randomized control trials (RCTs). However, RCTs are infeasible in personalized clinical risk prediction because the data collected in the trial may not be applicable to an individual's specific characteristics. In addition, due to the high dimensionality of clinical data collected in current Big Data technologies, applying conventional statistical analysis for causal analysis becomes incapable. Causal analysis has been applied in clinical knowledge mining. For example, authors in [16] applied causal analysis to assess whether the association of serum homocysteine concentration with ischaemic heart disease, deep vein thrombosis and pulmonary embolism, and stroke is causal and, if so, to quantify the effect of homocysteine reduction in preventing them.

As mentioned previously, applying causal analysis in clinical association rule mining is still rare even though the mined rules can still be potential indicators for the causality [17]. Therefore, it is important to combine other data mining technologies with association rules for causal analysis, which is important in the prediction of potential clinical risks based on a patient's individual characteristics.

2.3 Causality Metric for Association Rules

The antecedent of a rule can be called 'causal' regarding the consequent only if the antecedent is the trigger of the consequent, or the consequent is understood as a physical consequence of the antecedent. We can assess if a rule $X \Rightarrow Y$ has a high level of association according to its confidence value. However, high confidence of a rule $X \Rightarrow Y$ still cannot guarantee a low confidence of its counter case ($\bar{X} \Rightarrow Y$), which means that the observed conditions X may not be the real cause of the targeted outcome Y . During clinical decision support, the mining process should consider when the rule $X \Rightarrow Y$ yields a higher confidence than its antecedent's counter case (i.e., $\bar{X} \Rightarrow Y$). This means that Y is likely to occur only when X occurs. When X does not occur, Y has a low chance of occurrence. The following equation can be used to calculate the causality ratio of a rule $X \Rightarrow Y$:

$$\text{causality}(X \Rightarrow Y) = \frac{\text{confidence}(X \Rightarrow Y)}{\text{confidence}(\bar{X} \Rightarrow Y)}. \quad (3)$$

The causality ratio ranges from 0 to ∞ . The value of 1 is a significant threshold for the ratio. A rule with causality < 1 means that the antecedent predicts the consequent worse than the counter case of the antecedent, meaning that the causality of the outcome given the antecedent is low. This type of rule should be ignored, and all rules are expected to have causality ratios > 1 .

2.4 Causality-Based Rule Selection

Given a patient with N observed conditions with targeted predicted outcomes (i.e., the consequent), the mining framework can generate a list of up to $2^N - 1$ raw rules, regardless of the $\text{Supp}_{\min}$ and $\text{Conf}_{\min}$, in which the antecedent of each rule is a subset of the N observed conditions. Many methods have been proposed to determine the final prediction

possibility based on confidence values in the rule list. The Classification Based on Association (CBA) is the most common strategy that performs classification based on the selected rule's confidence values [18]. CBA uses a heuristic method to construct a rule list in which rules are ordered decreasingly according to their confidence and support values. Afterwards, the estimated possibility of the targeted outcome is based on the confidence value of the first rule (i.e., rule with highest confidence). For example, with a targeted outcome Y , a set of four observed clinical conditions $X = \{X_A, X_B, X_C, X_D\}$ from one patient can generate a list of $2^4 - 1 = 15$ rules in which the first rule R_I with antecedent events $\{X_A, X_B\}$ has the highest confidence of 90%. The CBA outputs this estimated possibility according to this rule. However, the list may contain other two rules R_{II} and R_{III} with antecedent events $\{X_A, X_C, X_D\}$ and $\{X_A, X_D\}$, which are both subsets of X , with confidence values of 30% and 23%, respectively. If all other rules in the list are similar to R_I and R_{III} that tend to have low confidence values, it may imply that the high confidence of R_I might happen by chance or mined from a few spurious data tuples. In this situation, the estimated 90% based on R_I in CBA becomes problematic. To improve the rule selection performance, we propose a new strategy that generates the final outcome prediction depending not only on rules' confidence values but also on their causality ratios calculated in Eq. (3). As illustrated in Figure 1, given a patient who has N observed conditions (X) with a fixed targeted outcome (Y), the mining framework can generate a list of $2^N - 1$ raw rules. The rule selection process first discards those rules with zero support values and causality ratios < 1 so as to guarantee that antecedents in all rules can be considered as potential causes with respect to the target outcome. Afterwards, the process generates two ranks on all remaining rules: one rank ($Rank_{CON}$) based on confidence values, and the other rank ($Rank_{CAU}$) based on causality ratios. A final rank is decided by the summation of $Rank_{CON}$ and $Rank_{CAU}$. With this final rank, the process calculates the average of confidence values from the top N_{Top} number of rules (i.e., $Rule_{SEL}$) and outputs it as the final prediction possibility.

3. PREDICTIVE RULE CALIBRATION

After generating a list with rules based on a patient's conditions, the rule selection process outputs estimated probability that the targeted outcome will occur (e.g., the patient has an 8% chance of having ICU mortality given her risk factors). This section discusses how to evaluate how close the estimated possibility to the real-world (i.e., observed) possibility.

The Area Under the ROC Curve (AUC), or c -statistic, has been widely used as a metric for discrimination model assessment, e.g., classification evaluation. Its popularity is increasing because it evaluates model performance across all possible cutoff values instead of an arbitrarily chosen cutoff [19]. However, the goal of this study is to estimate an actual probability of clinical events. Thus the AUC may not be the absolute best metric here because it only describes how well models can rank order cases and noncases, but is not a function of the actual predicted probabilities [3]. Alternatively, we can measure the calibration of the estimated prediction by checking how close this prediction is to the observed probability of targeted outcomes on a particular patient. If the prediction is close to the observed probability, the individualized estimate is considered 'well calibrated.'

Literature has supportive evidence that calibration is necessary in clinical decision support since outcome predictions are often used to determine a patient's individual risk [20–23].

A common measure of prediction calibration is Hosmer- Lemeshow's (HL) goodness-of-fit χ^2 -statistic, which compares observed and predicted outcomes for groups of risk. The HL test specifically identifies subgroups as the deciles of fitted risk values. Specifically, the predicted values are arrayed from lowest to highest, and then separated into several groups of approximately equal size. For each group, we calculate the observed number of outcomes and non-outcomes, as well as the expected number of outcomes and non-outcomes. The expected number of outcomes is just the sum of the predicted probabilities of the individuals in the group. And the expected number of non-outcomes is the group size minus the expected number of individuals without the outcome. Models are called well calibrated if expected and observed event rates in subgroups are similar. Therefore, the HL test statistic is given by:

$$HL_{\text{statistic}} = \sum_{j=1}^K \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi^2 \quad (4)$$

where O_j , E_j , and n_j denote the observed events, expected events, and number of observations in the j th group, and K is the number of groups. Pearson's chi-square (χ^2) is then applied to compare observed counts with expected counts. The HL statistic asymptotically follows a χ^2 distribution with $K - 2$ degrees of freedom (*dof*). The number of groups may be adjusted depending on how many fitted risks are determined by the model, which is important to avoid group with singular risks. The final calibrations can be determined based on the p -values that can be transformed from the HL statistic given the *dof*. A high p -value suggests a rejection of the model, meaning the model is well calibrated.

4. CASE STUDY

4.1 Background

Developing risk prediction models is one of the major purposes in the ICU [24]. Many illness scoring models are well developed and widely used, such as Acute Physiology and Chronic Health Evaluation (APACHE) [25], Logistic Organ Dysfunction Score (LODS) [26], Multiple Organ Dysfunction Score (MODS) [27], Mortality Prediction Model (MPM) [28], Simplified Acute Physiology Score (SAPS) [29], and Sequential Organ Failure Assessment (SOFA) [30]. Many recently developed models can handle more complicated clinical situations by applying advanced data mining approaches [31, 32]. However, a majority of these risk prediction models have common limitations, including relying on fixed attributes and fixed observation periods. Therefore, in our previous study, we proposed a temporal association rule mining framework, i.e., icuARM-II, that can estimate the development of potential clinical risks based on available patient conditions without being restricted by a fixed observation window [7]. In this case study, we adopt icuARM-II's personalized rule mining framework and evaluate the improvement of clinical risk predictions using our new causality- based rule selection strategy.

4.2 CHOA ICU Dataset

The dataset used in this case study was imported from the pediatric ICU database of Children's Healthcare of Atlanta (CHOA). This dataset contains clinical information collected in 5,739 ICU stays from 4,975 patients aged from birth to 21 years old during the year of 2013. The clinical information includes visit information, procedures, laboratory testing, and microbiology testing. Other than visit information, all data was temporal with timestamps, allowing us to perform temporal risk estimation using icuARM-II's mining framework. Examples and number of records in each category are tabularized in Table 1.

4.3 Lab results vs. six-hour ICU mortality

Using personalized laboratory (lab) testing results to develop ICU risk prediction models are relatively rare comparing to conventional scales [33], such as the admission-based APACHE- II [25] or the daily-based Sequential Organ Failure Assessment (SOFA) [30]. However, the current Big Data techniques promise a high volume of clinical attributes that can comprehensively cover individual characteristics so as to provide a more precise clinical risk prediction. In addition, to our best knowledge, there is no clinical decision support system that uses temporal and personalized lab testing abnormalities to predict upcoming clinical events. Therefore, in this case study, we applied icuARM-II to generate association rules for short-term (i.e., 6-hr) ICU mortality prediction based on personalized lab testing abnormalities.

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 frequent (i.e., counted) tests. The total number of records in each test and those in each level are listed in Table 2. To predict the 6-hr ICU mortality based on personalized lab results, temporal rules are in the form of $X \Rightarrow Y: \{E_A \mid T_A\} \Rightarrow \{Death \mid <6-hr\}$. All rules share the same targeted outcome, $\{Death \mid <6-hr\}$, predicting the *Death* event in the following six hours. Each rule has different antecedent episode X . Each X consists of a set of abnormal lab events E_A that have been observed in the past T_A time period. The generation of support and confidence values for each temporal association rules was based on the event scanning mechanism of the icuARM-II mining framework [7].

4.3.1 Method—To evaluate personalized predictions of the 6-hr ICU mortality, we generated 1,256 rules with random antecedents to simulate a variety of personalized clinical conditions. Each antecedent set consists of randomly selected lab abnormalities (i.e., E_A) chosen from the 12 items in Table 2. The antecedent episode represents a set of abnormal lab testing results E_A that have been observed in the past time period with a length of T_A to predict the *Death* event in the following six hours. Since abnormalities of a certain lab test could occur multiple times within an observation period, we simplified E_A with a chain of E^*N in which E was a lab testing item and N indicated its repeat. For instance, as the rule #7 in Table 3, if a patient has had abnormal potassium level twice and artery deficit once in the past 18-hr, the rule was represented as $\{POT^*2, ADE^*1 \mid <18-hr\} \Rightarrow \{Death \mid <6-hr\}$. In addition, the observation times (i.e., T_A) were randomly assigned to the 1,256 simulated rules, ranging from six hours to 48-hours. We selected ten examples to illustrate these rules in Table 3.

The flow of prediction evaluation is illustrated in Figure 2. The purpose of the evaluation was to investigate if the newly proposed (i.e., confidence-causality-based) rule selection (CC) strategy can achieve higher prediction calibration compared with the conventional top-confidence-based rule selection (TC). The flow of TC rule selection can also be depicted in Figure 1 without the block of $Rank_{AUC}$. We performed 5-fold ($K = 5$) cross validation to calculate the Hosmer-Lemeshow's (HL) χ^2 -statistic for the prediction calibration. For each fold, the $K-1$ groups of data was used as the training dataset for the generation of the 'estimated' predictions, $P_{e,CC}$ and $P_{e,TC}$, using the CC and TC, respectively. The remaining one group of data was used as the testing dataset for the generation of the 'observed' predictions, $P_{o,CC}$ and $P_{o,TC}$, using CC and TC, respectively. Since all groups were separated with equal sizes, the total observation in each group is known. The HL-statistic and the corresponding p-value (PV_{CC}) using CC can be calculated via $P_{e,CC}$ and $P_{o,CC}$. Similarly, the HL-statistic and the corresponding p-value (PV_{TC}) using TC can be calculated via $P_{e,TC}$ and $P_{o,TC}$. Afterwards, if $PV_{CC} > PV_{TC}$, we can verify that CC has better calibration than TC; otherwise, if $PV_{CC} < PV_{TC}$, CC has worse calibration than TC.

4.3.2 Results—We hypothesized that the prediction estimated by the newly proposed (i.e., confidence-causality-based) rule selection (CC) strategy had higher calibration compared with the conventional top-confidence-based rule selection (TC). Figure 3 shows the comparison results of prediction calibration using CC and TC on the 1,256 simulated personalized lab abnormality sets. As shown, 64.01% of condition sets were well calibrated by TC. Calibrations by CC depended on different number of selected top-ranked rules (N_{Top}), which can reach 89.49% when N_{Top} became significant. Regardless of N_{Top} , the CC has statistically and significantly ($p < 0.05$ via Student's t -test) better calibration than TC, which supports our hypothesis. Besides, CC can calibrate better than TC in 90.76% of cases when N_{Top} became significant. In Table 3, under the condition of $N_{Top} = 10$, we provide five rules with top $P_{e,CC}$ in the case of $PV_{CC} > PV_{TC}$ and another five rules with top $P_{e,CC}$ but $PV_{CC} < PV_{TC}$.

It is worth noting that all of the rules generated in this case study are examples to demonstrate the performance improvement of new rule selection strategy. In addition, we only used 12 lab testing items in this case study. The icuARM-II system 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. Based on a patient's instant conditions, clinicians can construct a personalized antecedent episode and a target consequent episode with flexible time windows.

5. CONCLUSION AND FUTURE WORK

Personalized risk prediction is becoming more important as the advance in clinical data collection techniques. However, risk prediction using temporal clinical data is still immature, and the assessment of the estimated evidence is still imperfect using conventional statistical analyses. In this paper, we improved the temporal rule mining framework developed in our previous icuARM-II system that can generate personalized risk prediction evidence using clinical data. We proposed and implemented a novel association rule selection mechanism combining a newly designed causality metric. The final risk prediction

possibilities were statistically assessed using a cross-validation-based calibration method. Finally, we conducted a case study to evaluate the performance improvement of the new rule selection mechanism. Our results showed an improved calibration in the prediction of short-term ICU mortality based on personalized lab testing abnormalities.

This work could be extended in three main directions. First, current electronic health record (EHR) system can comprehensively include a diverse health-related information from a patient. Advanced bedside monitoring technology can even collected intensive physiological data in the base of milliseconds. Therefore, other than the lab testing abnormalities that we used for method development and performance evaluation, we will expand our scope by including clinical variables from other categories, such as medical history, microbiology testing, and clinical procedures. This will allow us to provide a more precise clinical risk prediction. Second, we will apply the risk prediction to other clinical adverse outcomes (e.g., prolonged ICU stays and ventilator days, short readmission, and unnecessary care cost). Finally, interoperability is widely acknowledged as a key requirement for the success of healthcare information systems. Thus, we will deploy the current icuARM-II prototype on real-time EHR clinical system by adopting standardized healthcare information exchange guideline, such as the well-known FHIR (Fast Healthcare Interoperability Resources) specification [34].

Acknowledgments

We thank Matthew Miller, Michael Thompson, Sherry Farrugia, and Tod Davis for helping the development of the CHOA ICU database. The authors are also grateful to Dr. Nikhil Chanani and Dr. Kevin Maher for providing clinical significances when building the icuARM-II system that are used in the case study study. The authors also thank Ryan Hoffman and Tong Li for their valuable comments and suggestions. This research has been supported by grants from NIH (R01CA163256 and R01CA163256), CHOA Center for Pediatric Intervention (CPI), Georgia Cancer Coalition Award to Prof. MD Wang, Hewlett Packard, and Microsoft Research.

REFERENCES

1. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. *Statistics in medicine*. 2011; 30(10):1090–1104. [PubMed: 21337591]
2. Jiang X, Boxwala AA, El-Kareh R, Kim J, Ohno-Machado L. A patient-driven adaptive prediction technique to improve personalized risk estimation for clinical decision support. *Journal of the American Medical Informatics Association*. 2012 amiajnl-2011-000751.
3. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007; 115(7):928–935. [PubMed: 17309939]
4. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical chemistry*. 2008; 54(1):17–23. [PubMed: 18024533]
5. Sacchi L, Larizza C, Combi C, Bellazzi R. Data mining with temporal abstractions: learning rules from time series. *Data Mining and Knowledge Discovery*. 2007; 15(2):217–247.
6. Gharib TF, Nassar H, Taha M, Abraham A. An efficient algorithm for incremental mining of temporal association rules. *Data & Knowledge Engineering*. 2010; 69(8):800–815.
7. Cheng, C-W.; Chanani, N.; Maher, K.; Wang, MD. icuARM-II: improving the reliability of personalized risk prediction in pediatric intensive care units; *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*; 2014. p. 211–219.
8. Agrawal R, Imieli ski T, Swami A. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*. 1993; 22(1):207–216.
9. Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining—a general survey and comparison. *ACM SIGKDD Explorations Newsletter*. 2000; 2(1):58–64.

10. Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, VLDB. 1994; 1215:487–499.
11. Konias S, Giaglis G, Gogou G, Bamidis P, Maglaveras N. Uncertainty rule generation on a home care database of heart failure patients. Computers in Cardiology, 2003. 2003:765–768.
12. Shan, Y.; Jeacocke, D.; Murray, DW.; Sutinen, A. Mining medical specialist billing patterns for health service management; Proceedings of the 7th Australasian Data Mining Conference-Volume 87; 2008. p. 105–110.
13. 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]
14. 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]
15. Chang C-L. A study of applying data mining to early intervention for developmentally-delayed children. Expert Systems with Applications. 2007; 33(2):407–412.
16. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta- analysis. Bmj. 2002; 325(7374):1202. [PubMed: 12446535]
17. Hill AB. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965; 58(5):295. [PubMed: 14283879]
18. Ma, BLWHY. Integrating classification and association rule mining; Proceedings of the fourth international conference on knowledge discovery and data mining; 1998.
19. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 1997; 30(7):1145–1159.
20. Karp I, Abrahamowicz M, Bartlett G, Pilote L. Updated risk factor values and the ability of the multivariable risk score to predict coronary heart disease. American journal of epidemiology. 2004; 160(7):707–716. [PubMed: 15383415]
21. Wei W, Visweswaran S, Cooper GF. The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. Journal of the American Medical Informatics Association. 2011; 18(4):370–375. [PubMed: 21672907]
22. Ohno-Machado L, Resnic FS, Matheny ME. Prognosis in critical care. Annu. Rev. Biomed. Eng. 2006; 8:567–599. [PubMed: 16834567]
23. 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]
24. Rosenberg AL. Recent innovations in intensive care unit risk-prediction models. Current opinion in critical care. 2002; 8(4):321–330. [PubMed: 12386493]
25. 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]
26. Le Gall J-R, 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]
27. 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]
28. 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]
29. Le Gall J-R, 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]
30. Vincent J-L, Ferreira F, Moreno R. Scoring systems for assessing organ dysfunction and survival. Critical care clinics. 2000; 16(2):353–366. [PubMed: 10768086]

31. 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.
32. Clark PA, Lettieri CJ. Clinical model for predicting prolonged mechanical ventilation. *Journal of critical care*. 2013; 28(5):880, e881-880, e887. [PubMed: 23683556]
33. 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]
34. Bender, D.; Sartipi, K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange; *Computer-Based Medical Systems (CBMS)*, 2013 IEEE 26th International Symposium on; 2013. p. 326-331.

Figure 1.

Flow of causality-based rule selection process. The process selects rules based on the ranks of confidence and causality values of all rules generated from personalized clinical conditions.

Figure 2.

Flow of prediction assessment using newly proposed (i.e., confidence-causality-based) rule selection (CAU) strategy and conventional top-confidence-based rule selection (TC).

Figure 3.

Calibration comparison results between the newly proposed confidence-causality-based rules selection (CC) and the conventional top-confidence rule selection (TC). (a) Percentage of CC and TC that are well calibrated. CC with different number of top-ranked rules all calibrate better than TC. (b) The percentage that CC calibrates better than TC with different number of top-ranked rules.

Table 1

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 2

Top 12 counted lab testing items in CHOAICU database

Lab Item	# of Records		
	Total	Normal	Abnormal
Glucose	58,621	19,793 (33.8%)	38,828 (66.2%)
Arterial PCO ₂	49,246	26,898 (54.6%)	22,348 (45.4%)
Arterial pH	49,246	21,772 (44.2%)	27,474 (55.8%)
Arterial PO ₂	49,246	5,733 (11.6%)	43,513 (88.4%)
Ox Saturation	49,246	22,035 (44.7%)	27,211 (55.3%)
Ionized Ca	43,174	28,690 (66.5%)	14,484 (33.5%)
Potassium	40,925	24,232 (59.2%)	16,693 (40.8%)
Sodium	39,724	24,870 (62.6%)	14,854 (37.4%)
Total CO ₂	49,246	42,898 (87.1%)	6,348 (12.9%)
Art. Excess	25,700	11,198 (43.6%)	14,502 (56.4%)
Art. Deficit	23,595	8,533 (36.2%)	15,062 (63.8%)
Creatinine	21,870	18,607 (85.1%)	3,263 (14.9%)
Total	499,839	255,259 (51.1%)	244,580 (48.9%)

Table 3

Example Personalized Temporal Association Rules Sorted by $P_{e,CC}$ with 20 top selected rules ($N_{top} = 10$)

#	Antecedent		Consequent		CC		TC	
	Observed Events (E_A)	Observed Time (T_A)	Predicted Event	Predicted Time	$P_{e,CC}$	PV_{CC}	$P_{e,TC}$	PV_{TC}
$PV_{CC} > PV_{TC}$	1 APCO2*1 + APH*5 + APO2*3	12hr	<i>Death</i>	6-hr	14.85%	0.76	15.32%	0.23
	2 APCO2*1 + APH*3 + APO2*3	24hr			13.82%	0.84	14.23%	0.03
	3 APCO2*1 + ICA*4 + APH*4	18hr			13.13%	0.53	14.46%	0.47
	4 APCO2*1 + APO2*2 + ICA*2	24hr			12.24%	0.68	13.18%	2e-4
	5 AEX*1+ICA*1+CRE*2	12hr			12.13%	0.77	12.58%	0.02
$PV_{CC} < PV_{TC}$	6 APCO2*1 + POT*1 + ADE*1	18hr			10.31%	3e-3	11.32%	0.13
	7 POT*2 + ADE*1	18hr			9.85%	0.24	10.21%	0.34
	8 CRE*1+ADE*2	12hr			9.83%	1e-4	10.23%	0.08
	9 APO2*1+ICA*2	18hr			9.24%	0.21	9.79%	0.43
	10 SOD*1+CO2*2	24hr			9.13%	2e-3	9.45%	0.11

ADE: Art. Deficit APCO2; Arterial PCO2 APH: Arterial pH APO2: Arterial PO2 AXE: Art. Excess CO2: Total CO2 CRE: Creatinine ICA: Ionized Ca POT: Potassium SOD: Sodium

Note 1: #1 to #5 are rules with $PV_{CC} > PV_{TC}$ (i.e., CC is better calibrated than TC)

Note 2: #6 to #10 are rules with $PV_{CC} < PV_{TC}$ (i.e., TC is better calibrated than CC)