Analyzing Electronic Medical Records to Predict Risk of DIT (Death, Intubation, or Transfer to ICU) in Pediatric Respiratory Failure or Related Conditions

Teeradache Viangteeravat, PhD[1,2], Oguz Akbilgic, PhD[2,3,4], Robert Lowell Davis, MD, MPH[2,3]

[1]Biomedical Informatics Core, Children's Foundation Research Institute, Le Bonheur Children's Hospital, Memphis, TN; [2]Department of Pediatrics, [3]UTHSC-ORNL Center for Biomedical Informatics, [4]Department of Preventive Medicine, The University of Tennessee Health Science Center, Memphis, TN, USA

Abstract

Large volumes of data are generated in hospital settings, including clinical and physiological data generated during the course of patient care. Our goal, as proof of concept, was to identify early clinical factors or traits useful for predicting the outcome, of death, intubation, or transfer to ICU, for children with pediatric respiratory failure. We implemented both supervised and unsupervised methods to extend our understanding on statistical relationships in clinical and physiological data. As a supervised learning method, we use binary logistic regression to predict the risk of developing DIT outcome. Next, we implemented unsupervised k-means algorithm on principal components of clinical and physiological data to further explore the contribution of clinical and physiological data on developing DIT outcome. Our results show that early signals of DIT can be detected in physiological data, and two risk factors, blood pressure and oxygen level, are the most important determinant of developing DIT.

Introduction

In the healthcare industry, a considerable amount of data is generated through routine clinical practice, including data for patient blood pressure, oxygen level, respiratory rate, heart rate, hemoglobin and hematocrit, and blood electrolytes. While these data are used primarily to monitor patient clinical status, their use in the field of predictive analytics – for the purpose of predicting upcoming clinical deterioration - has been relatively limited. Healthcare institutions are increasingly joining the ranks of other major industries in using numerous data mining techniques to identify trends and hidden relationships in these large and complex data sets[1, 2]. Structured record data has been used to extract phenotype information from free-text records to produce fine-grained disease correlations and patient stratification[3]. Such data has also been used to create similarity ranking (or matrices) between pairs of patients with rare diseases[4]. Data mining techniques have been widely employed in the fields of clinical informatics and genomics[5]. Bayesian models are among the most widely applied techniques in medical applications, and are used to classify data into supervised learning classes[6]. For free-text documents, latent semantic indexing has been widely used to produce a concept vector space in which query vectors and term-document are projected[7]. For example, latent semantic indexing techniques have facilitated the extraction of gene function data from peer-reviewed scientific abstracts, which contributes to the understanding of high-throughput genomic studies[8].

Mining data from medical information systems has been useful for predictive modeling, for example, to guide preventive care or to inform the healthcare team of critical patient signs indicative of disease or acute clinical crises. Symbolic time series approaches have been used to study physiological data; using symbolic time series analysis techniques, heart rate variability dynamics have been shown to distinguish healthy subjects from patients with cardiac problems[9]. Using features extracted from symbolic series and time-frequency indices of heart rate variability, Aziz *et al.* (2004) suggests that the use of new features based on symbolic series, coupled with classic time-frequency and clinical indices, is a good predictor of death in patient with Chagas disease. Symbolic time series analyses have also been applied to heart period (RR) and QT variability, and can improve separation between ischemic dilated cardiomyopathy patients and a healthy control group[10].

Current research has successfully developed an early detection signaling system capable of identifying young adolescents at high risk for sepsis[11], and has also allowed for the rapid identification of patients with possible septic shock in order to enroll them into a time sensitive clinical study[12]. Other examples of successful applications of data mining techniques to complex healthcare data include its use to guide hypertension management[13], and to identify

factors contributing to preterm birth[14]. Previously, we studied pediatric asthma patients by mining data from electronic medical records[15] using low-rank matrix decomposition (LRMD) in vector space models[16]. LRMD techniques were applied to the parse All Patient Refined Diagnosis Related Group (APR-DRG) datasets for asthma, allowing for the extraction of dominant features and the prediction of outcomes.

Asthma and acute lower respiratory tract infections are the single most common causes of hospitalization annually at Le Bonheur Children's Hospital (LBCH; a large referral hospital in Memphis, TN), and account for the majority of hospitalization during the winter months[17]. Children hospitalized with asthma or lower respiratory tract infections that have persistent episodes of hypoxemia (or require increasing fraction of inspired oxygen (FiO_2)) are more likely to require ICU transfer or need mechanical ventilation. A system that enables early recognition of declining respiratory function could affect real change in a clinical setting, and in turn may help improve medical outcomes.

Methods

Setting & Participants

We used data from all patients admitted to LBCH with a diagnosis of asthma or related pulmonary conditions (such as wheezing and bronchiolitis) from January 2013 to April 2013. All data used for this study was extracted from the LBCH 'Cerner' Electronic Medical Records. The total number of observations included in our study includes 745 encounters from 563 distinct patients. Of these 563 patients, 60% were African American, 28% Caucasian, <1% were Asian American, and 12% were 'other' race; 53% were male, and 28% were between 1 and 4 years old. Of the study cohort, 10.5% (n=59) required ICU transfer, approximately 2% (n=11) required mechanical ventilation and 0.1% (n=1) died. The UTHSC Institutional Review Board (IRB) approved this study for exempt status.

Variable Selection & Model Building Process

In data mining, the selection of the set of explanatory variables (or predictors) is typically part of the analysis. For our approach, we used an automatic variable selection procedure, stepwise regression based on Akaike Information Criteria[18, 19]. We used Beta or standardized coefficients after converting all variables to z-scores prior to the variable selection process. Standardized coefficients allow a comparison of the relative importance of the risk or predictor variables. Since the outcome of 'death, intubation, or transfer to the ICU' was dichotomous in nature (DIT=No, DIT=Yes), we used binary logistic regression to support the evaluation of multiple risk factors[20, 21].

Unlike age and gender, some variables such as FiO_2, SpO_2, BP, MCV and respiratory rate consist of multiple measurements over time for each patient. In order to handle the multiple measurements obtained within one hospitalization for these potential risk factors, we converted the multiple measurements into a single value, based on the selection criteria outlined in Table 1. This approach was used to minimize data loss associated with converting to descriptive statistics (for example, mean or median). To study the time window effect on DIT, we used Table 1 in processing variables with multiple measurements in two different ways; 48 and 12 hours prior to DIT. Therefore, we created two separate logistic predicting the risk of developing DIT using the variables in Table 1. This allows us to explore both whether there were different variables associated with DIT at different time windows before the occurrence of DIT and whether the variables associated with DIT had different strengths of association when measured in these different time windows prior to DIT.

Table 1. Potential Risk Factors

Potential Risk Factor	**Selection Criteria**
Fraction of Inspired Oxygen (FiO_2)†	Number of times $FiO_2 > 0.5$
Oxygen Saturation (SpO_2) ‡	Number of times $SpO_2 < 90$
Mean Corpuscular Volume of Blood Cell (MCV)	First value measured after admission
Mean Corpuscular Hemoglobin Concentration in blood (MCHC)	First value measured after admission
Respiratory rate§	Number of times respiratory rate less than or above normal age-specific range
Blood pressure (Systolic)§	Number of times blood pressure less than or above the normal age-specific range

Table 1. Potential Risk Factors (continued)

Sodium	First value measured after admission
Potassium	First value measured after admission
Gender	Male/Female
Race	African American, White, Asian, others
Age	Between 1 and 18 years old

† = FiO_2 is typically maintained below 0.5 even with mechanical ventilation (to avoid oxygen toxicity); ‡ = Normal pulse oximeter readings usually range from 95 to 100 percent. SpO_2 values under 90 percent are considered low and usually indicate the need for supplemental oxygen; § = we used the standard primary vital signs that are provided by American College of Emergency Physicians[22]

Further, we applied principal component analysis (PCA) on explanatory variables to reduce the dimension and to examine which clinical factors are most strongly correlated with each principal component. In our case, a correlation value above 0.5 magnitudes in either positive or negative direction is deemed important. Next, we carried out *k*-means clustering algorithm[23] on selected principle components to find patterns between the risk factors and DIT.

Results

Our logistic regression analysis results for the data organized for 48 hours prior to DIT are shown in Table 2. Our results suggest that among all variables collected, FiO_2 is most strongly associated with the outcome, followed by MCV, respiratory rate, and the subject's race (Table 2). Figure 1 shows the receiver operating curve for our final model, presenting the true positive rate versus false positive rate; the area under the curve (AUC; the c-index or c-statistic) is 0.875.

Table 2. Multivariable logistic regression (48 hours prior to DIT outcomes).

					95% confidence interval	
	Parameter Estimate	Standard Error	*p* Value	Odds Ratio	CI Lower Limit	CI Upper Limit
Age	-0.026	0.082	0.752	0.974	0.825	1.149
Gender						
Male	0.541	0.471	0.251	1.716	0.687	4.407
Race*						
African American	1.492	0.679	**0.028**	4.445	1.245	18.463
Other	2.089	0.883	**0.018**	8.078	1.478	4.957
SpO_2	-0.025	0.027	0.347	0.975	0.921	1.029
BP systolic	0.056	0.034	0.096	1.057	0.993	1.134
Respiratory rate*	0.057	0.022	**0.008**	1.059	1.016	1.108
MCHC	-0.219	0.199	0.271	0.803	0.534	1.176
MCV*	0.099	0.034	**0.003**	1.105	1.038	1.186
Potassium	-0.082	0.357	0.819	0.922	0.451	1.858
Sodium	-0.112	0.068	0.098	0.894	0.776	1.019
FiO_2*	0.257	0.074	< **0.001**	1.293	1.145	1.529

The five variables most strongly associated with Death, Intubation and ICU Transfer are shown in bold. * = The odds ratios for these variables (with 95% confidence intervals) are significant at the 0.05 level.

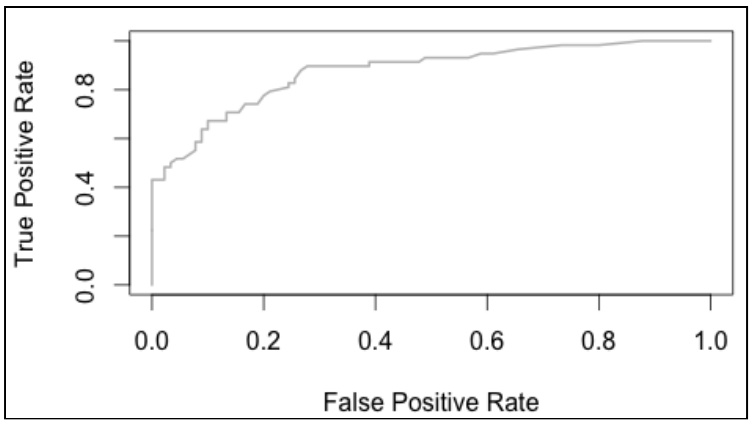

Figure 1. Receiver Operating Curve (true positive vs. false positive).

Our second logistic regression model using the data organized for 12 hours prior to DIT is shown in Table 3. Among variables measured 12 hours prior to DIT, blood pressure was significantly associated with DIT, along with race and FiO_2, but MCV and respiratory rate were not found to be strongly associated with DIT at this time point.

Table. 3 Multivariable logistic regression (less than 12 hours prior to DIT outcomes)

					95% confidence interval	
	Parameter Estimate	Standard Error	*p* Value	Odds Ratio	CI Lower Limit	CI Upper Limit
Age	0.045	0.127	0.724	1.046	0.821	1.357
Gender						
Male	0.055	0.785	0.944	1.056	0.209	4.939
Race*						
African American	3.224	1.153	**0.005**	25.143	3.168	318.144
Other	1.974	1.527	0.196	7.205	0.409	182.855
SpO_2	0.001	0.007	0.945	1.001	0.987	1.013
BP systolic*	0.029	0.006	< **0.001**	1.029	1.019	1.044
Respiratory rate	-0.004	0.006	0.556	0.996	0.983	1.008
MCHC	-0.027	0.314	0.930	0.973	0.524	1.809
MCV	0.448	0.049	0.367	1.046	0.951	1.158
Potassium	0.101	0.676	0.882	1.106	0.288	4.327
Sodium	-0.233	0.144	0.106	0.792	0.583	1.029
FiO_2*	0.549	0.171	**0.001**	1.732	1.283	2.522

The three variables most strongly associated with Death, Intubation and ICU Transfer are shown in bold. * = The odds ratios for these variables (with 95% confidence intervals) are significant at the 0.05 level.

We applied a PCA technique [24] to simplify our data set into a lower dimensional space to allow for visualization of associations in the data. Figure 2 shows a plot of variances (y-axis) associated with principal components (x-axis). We selected the first 3 components for our analysis, based on the variability in the data using the "elbow" method of scree plot (Figure 2)

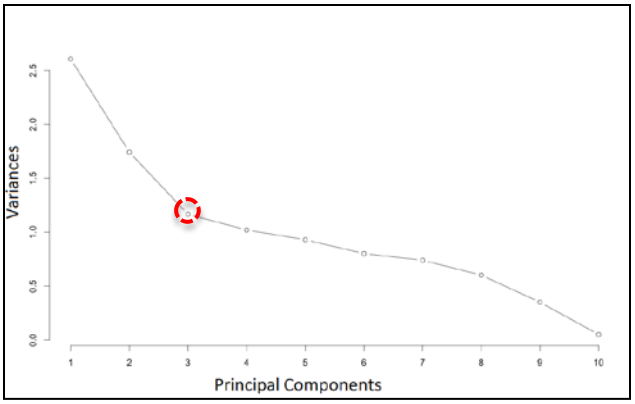

Figure 2. The plot of variances (y-axis) that is associated with each principal component using PCA. The "elbow" is shown by the red circle.

A correlation value of magnitude 0.5 was assumed as significant. We found that the first principal component is highly correlated with BP systolic (correlation co-efficient of -0.547), while SpO_2 (0.565) is correlated with the second component. Age (-0.686) and MCV (-0.596) are correlated with the third principal component. To examine patterns in the data with respect to DIT, we applied an unsupervised machine learning technique, *k*-means clustering analysis [23], using the first three principal components. To determine the number of clusters, we looked at the within groups sum of squares and selected the "elbow" in the plot. We can see that the "elbow" in the scree plot is at $k = 3$ (Figure 3) so we applied the *k*-mean clustering with $k = 3$.

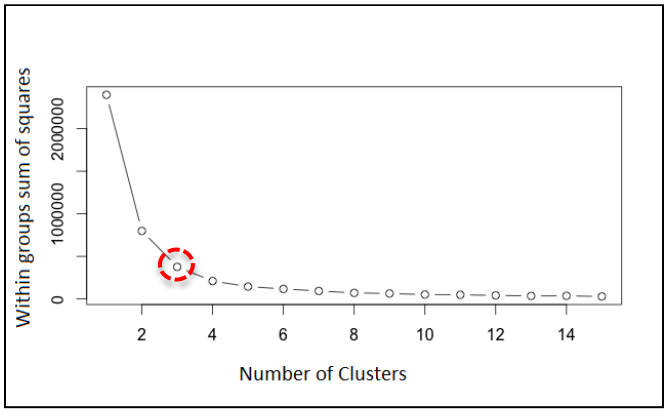

Figure 3. Number of clusters vs. within groups sum of squares using the "nstart = 25" and "iter.max = 1000" in R version 3.2.3. The "elbow" is shown by the red circle.

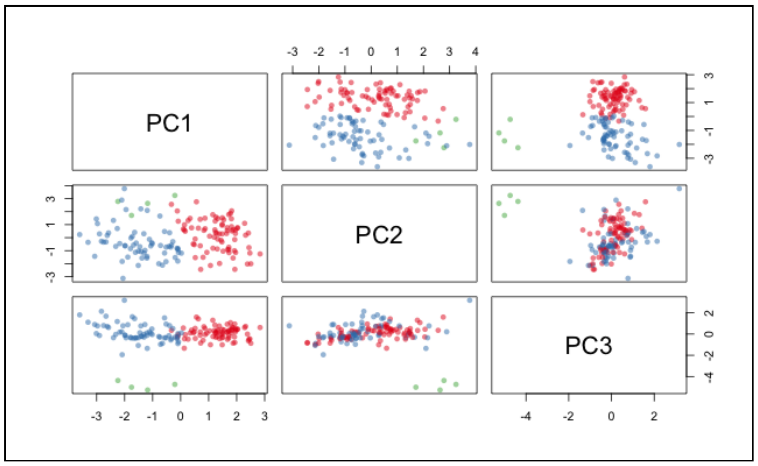

Figure 4. Applied *k*-mean clustering with $k = 3$ to the first three principal components (PC1, PC2, and PC3).

We then applied the *k*-mean clustering with $k = 3$ using the first three principal components (Figure 4). In each box, principal components are compared and are either comingled (mixture of blue and red dots) or clustered separately (blue cluster and red cluster). Outliers can be clearly identified, and are shown in green. Figure 4 shows separation between PC1 and PC2 clusters, therefore we focused on BP systolic and SpO_2 as we previously identified these variables as highly correlated with the first two principal components. Despite this result, SpO_2 was not one of the statistically significant predictors of DIT in our logistic regression models in Table 2 and Table 3. Also, BP systolic was significant when we analyzed our data for 12 hours prior to DIT, but not at 48 hours prior to DIT (see Table 3). Therefore, we studied these two variables' associations to DIT by projecting the *k*-mean clustering results with $k = 3$ on these two variables (Figure 5).

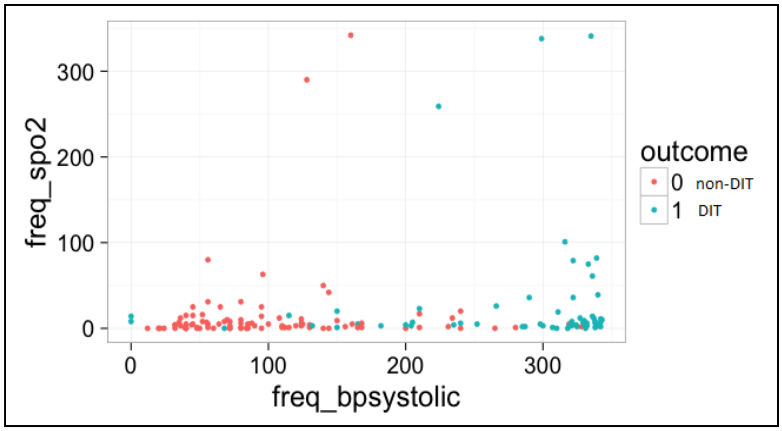

Figure 5. Count of the number of times that BP systolic values reach less than normal values or above normal values (adjusted for the patient's age) (freq_bpsystolic) versus the number of times SpO_2 values reach less than 90 (freq_spo2). 0 = patient without DIT outcomes (red dot); 1 = patient with DIT outcomes (blue dot).

Plotting abnormal fluctuations in BP systolic versus SpO_2 revealed clustering of DIT (blue dot) and non-DIT (red dot) based on BP systolic, indicating that BP systolic is a promising variable for predicting our outcome. We did not observe a similar clustering pattern for SpO_2. To visualize the relationship between three clusters (from Figure 3) and the variables BP systolic and SpO_2, we used a *k*-mean algorithm to project BP systolic and SpO_2 variables onto three clusters (Figure 6). The count of the number of times BP systolic was less than or above the normal values is the main indicators assigning cases into these three clusters (Figure 6). The result of *k*-mean clustering analysis applied to BP systolic yielded promising results among patients who later progressed to DIT (i.e., transitioning from Ck_1 to Ck_2) and distinguished them from patients who remained healthy (i.e., transitioning from Ck_1 to Ck_3).

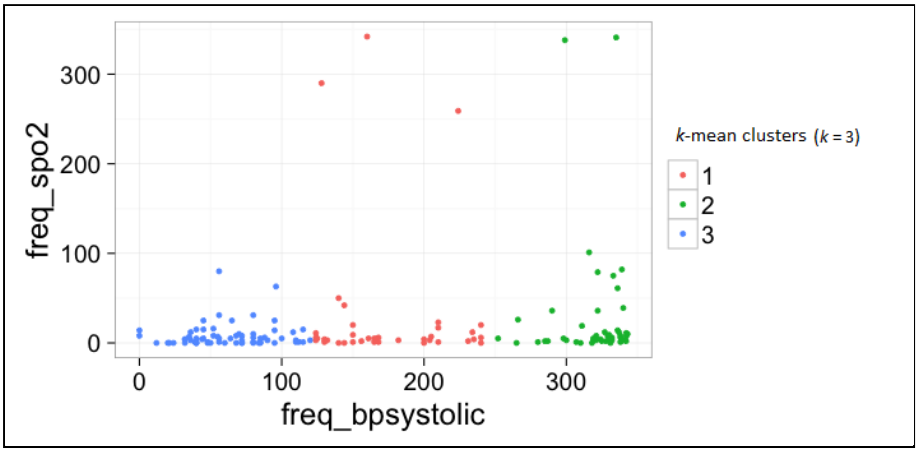

Figure 6. *k*-mean clustering results for $k = 3$; 1 = cluster number 1 (blue dot) with the centroid of "freq_bpsystolic" = 56.35 (or $Ck_1 = 56.35$); 2 = cluster number 2 (red dot) with the centroid of "freq_bpsystolic" = 151.51 (or $Ck_2 = 151.51$); 3 = cluster number 3 (green dot) with the centroid of "freq_bpsystolic" = 309.52 (or $Ck_3 = 309.52$); "freq_bpsystolic" = the number of times BP systolic is less than or above the normal values (adjusted for patient age).

Discussion

In this study, we found that race, respiratory rate, MCV and FiO_2 were significantly associated with impending respiratory failure - defined in this study as intubation, transfer to the ICU, or death. Interestingly, the strongest predictor of impending DIT is MCV. An abnormal MCV was associated with risk for respiratory deterioration 48 hours prior to DIT. MCV, along with MCH and MCHC, are part of red blood cell count indices and represent size, content, and hemoglobin concentration, but the biologic explanation for its association with DIT is not clear. In addition, we found that systolic BP was a key clinical factor predictive of DIT, especially in the time window 12 hours prior to DIT.

We found some evidence that predictive risk factors differ in their association with respiratory failure depending on the timing of these events. For example, abnormalities in some risk factors may be detected 48 hours before DIT, such as respiratory rate, MCV and FiO_2, while abnormalities in BP readings are significant for DIT in the following 12 hours. Since some of these traits, such as blood pressure, may also be evaluated in the context of a series of observations in a specific time interval, our future work will investigate the optimal time lag to be used for characterizing pattern transition behaviors like a Markov chain or Hidden Markov Model (HMM) to further identify specific patterns of time series among patients at particularly high risk for DIT. While our study did not use information downloaded directly from the cardiorespiratory monitors for time series data and these values were 2-3 hourly averages calculated and entered by the bedside nurse in the Electronic Medical Records, we believe that our analyses are suitable to apply to information downloaded directly from the cardiorespiratory monitors, which will allow for real-time monitoring in the future.

Study Limitations

One challenge to predictive modeling is that of generalizability. Because our study was relatively small, we elected against using split samples to validate our models. Therefore, our future plans will focus on verifying that our predictive models are robust and generalizable beyond just the data in hand. We will apply our predictive model to larger datasets such as the Pediatric Health Information System (PHIS) database, which consists of data from 44 leading children's hospital.

Conclusion

Analyzing Electronic Medical Records by applying data mining and clustering analysis techniques facilitates the possibility of discovering unexpected relationships and trends to gain new insights. Indeed, we hope that results from this study will advance progress toward the goal of identifying patients at high risk for respiratory failure. We believe our techniques can be expanded to encompass other diseases amendable to such modeling.

Acknowledgments

The authors would like to thank the University of Tennessee Health Science Center (UTHSC) Department of Information Technology Services Computing Systems division and the Center for Biomedical Informatics for the informatics resources and collaboration opportunities they offered. The author gratefully acknowledges Amanda Preston for editing and providing comments. The Children's Foundation Research Institute (CFRI), Le Bonheur Children's Hospital, and UTHSC supported this work. The authors have no competing interests to declare.

Author Contributions

TV conceived and drafted the manuscript with contributions from BD and OA.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Poulymenopoulou, M.; Malamateniou, F.; Vassilacopoulos, G. Machine Learning for Knowledge Extraction from PHR Big Data, *Studies in health technology and informatics*. **2014, *202,*** 36-39.
2. Yoo, C.; Ramirez, L.; Liuzzi, J. Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine, *Int Neurourol J*. **2014, *18,*** 50-57.
3. Roque, F. S.; Jensen, P. B.; Schmock, H., et al. Using electronic patient records to discover disease correlations and stratify patient cohorts, *PLoS computational biology*. **2011, *7,*** e1002141.
4. Davis, D. A.; Chawla, N. V.; Christakis, N. A.; Barabasi, A. L. Time to CARE: a collaborative engine for practical disease prediction, *Data Min Knowl Disc*. **2010, *20,*** 388-415.

5. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of machine learning*, Cambridge, MA: MIT Press, 2012.
6. Gelman, A. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing, *Int Stat Rev*. **2003, *71,*** 369-382.
7. Available at: http://www.puffinwarellc.com/index.php/news-and-articles/articles/33-latent-semantic-analysis-tutorial.html?showall=1. Accessed January 5, 2016.
8. Homayouni, R.; Heinrich, K.; Wei, L.; Berry, M. W. Gene clustering by latent semantic indexing of MEDLINE abstracts, *Bioinformatics*. **2005, *21,*** 104-115.
9. Aziz, W.; Rafique, M.; Ahmad, I.; Arif, M.; Habib, N.; Nadeem, M. S. Classification of heart rate signals of healthy and pathological subjects using threshold based symbolic entropy, *Acta biologica Hungarica*. **2014, *65,*** 252-264.
10. Valencia, J. F.; Vallverdu, M.; Rivero, I., et al. Symbolic dynamics to discriminate healthy and ischaemic dilated cardiomyopathy populations: an application to the variability of heart period and QT interval, *Philosophical transactions Series A, Mathematical, physical, and engineering sciences*. **2015, *373***.
11. Sepanski, R. J.; Godambe, S. A.; Mangum, C. D.; Bovat, C. S.; Zaritsky, A. L.; Shah, S. H. Designing a pediatric severe sepsis screening tool, *Frontiers in pediatrics*. **2014, *2,*** 56.
12. Herasevich, V.; Pieper, M. S.; Pulido, J.; Gajic, O. Enrollment into a time sensitive clinical study in the critical care setting: results from computerized septic shock sniffer implementation, *J Am Med Inform Assn*. **2011, *18,*** 639-644.
13. Chae, Y. M.; Ho, S. H.; Cho, K. W.; Lee, D. H.; Ji, S. H. Data mining approach to policy analysis in a health insurance domain, *International journal of medical informatics*. **2001, *62,*** 103-111.
14. Prather, J. C.; Lobach, D. F.; Goodwin, L. K.; Hales, J. W.; Hage, M. L.; Hammond, W. E. Medical data mining: knowledge discovery in a clinical data warehouse, *Proceedings : a conference of the American Medical Informatics Association / AMIA Annual Fall Symposium AMIA Fall Symposium*. **1997,** 101-105.
15. Viangteeravat, T.; Nagisetty, N. S. Giving raw data a chance to talk: a demonstration of exploratory visual analytics with a pediatric research database using Microsoft Live Labs Pivot to promote cohort discovery, research, and quality assessment, *Perspectives in health information management / AHIMA, American Health Information Management Association*. **2014, *11,*** 1d.
16. Viangteeravat, T. Potential identification of pediatric asthma patients within pediatric research database using low rank matrix decomposition, *Journal of clinical bioinformatics*. **2013, *3,*** 16.
17. CHAMP program. Retrieved from http://www.lebonheur.org/kids-health-wellness/le-bonheur-in-the-community/champ/.
18. Peduzzi, P. N.; Hardy, R. J.; Holford, T. R. A stepwise variable selection procedure for nonlinear regression models, *Biometrics*. **1980, *36,*** 511-516.
19. Arunajadai, S. G. Stepwise logistic regression, *Anesthesia and analgesia*. **2009, *109,*** 285; author reply 285-286.
20. Anderson, R. P.; Jin, R.; Grunkemeier, G. L. Understanding logistic regression analysis in clinical reports: an introduction, *The Annals of thoracic surgery*. **2003, *75,*** 753-757.
21. Sperandei, S. Understanding logistic regression analysis, *Biochemia medica*. **2014, *24,*** 12-18.
22. American College of Emergency Physicians. ER 101: Vital Signs. American College of Emergency Physicians. Available at http://www.emergencycareforyou.org/VitalCareMagazine/ER101/Default.aspx?id=500. Accessed: December 11, 2014.
23. Bishop, C. M. *Neural networks for pattern recognition*, Oxford New York: Clarendon Press ;Oxford University Press, 1995.
24. Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments, *Philos T R Soc A*. **2016, *374***.