Leveraging Derived Data Elements in Data Analytic Models for Understanding and Predicting Hospital Readmissions

Sharath Cholleti, PhD[1], Andrew Post, MD, PhD[1], Jingjing Gao, PhD[1], Xia Lin, PhD[1], William Bornstein, MD, PhD[2], Dedra Cantrell, RN[3], Joel Saltz, MD, PhD[1]
[1]Department of Biomedical Informatics, Emory University, Atlanta, GA
[2]Office of Quality, Emory Healthcare, Atlanta, GA
[3]Department of Information Services, Emory Healthcare, Atlanta, GA

Abstract

Hospital readmissions depend on numerous factors. Automated risk calculation using electronic health record (EHR) data could allow targeting care to prevent them. EHRs usually are incomplete with respect to data relevant to readmissions prediction. Lack of standard data representations in EHRs restricts generalizability of predictive models. We propose developing such models by first generating derived variables that characterize clinical phenotype. This reduces the number of variables, reduces noise, introduces clinical knowledge into model building, and abstracts away the underlying data representation, thus facilitating use of standard data mining algorithms. We combined this pre-processing step with a random forest algorithm to compute risk for readmission within 30 days for patients in ten disease categories. Results were promising for encounters that our algorithm assigned very high or very low risk. Assigning patients to either of these two risk groups could be of value to patient care teams aiming to prevent readmissions.

Introduction

Reducing the rate of hospital readmissions within 30 days has become a major focus of hospital quality improvement efforts due to upcoming payment incentives and penalties[1]. Previous efforts to predict readmissions tend to perform poorly and do not generalize well to different diseases, hospitals or regions[2]. The scope of data captured by hospital EHRs is growing, thus in theory making prediction easier. However, social history characteristics that are important factors in readmissions[2] still are seldom present in structured form, and in general hospitals differ in data representation and application even of standard coding systems. Time course of patient illness may be important in predicting future adverse events such as readmissions, and standard data mining algorithms have limited support for detecting such relationships. Similarly, meaningful groupings of codes may be hard for such algorithms to find because there likely are few examples of patients with a particular combination of codes even in large EHR datasets. Abstracting away these hospital-specific, temporal and high-dimensionality features and instead building models in terms of higher-level patient phenotype categories may allow creation of more accurate models.

Models of hospital readmission within 30 days potentially have broad applicability across the healthcare enterprise. As a preliminary step, testing strength of association of phenotypes with 30-day readmissions using retrospective data may help determine the applicability of a phenotype for prediction. Predictive models that incorporate phenotypes found to be of potential use may be tested retrospectively, withholding data from the model to simulate its application in clinical care. A predictive model with good accuracy could be implemented for real-time risk computation during a hospitalization. Upon admission, such a model could provide information for determining the intensity of care and what services to provide. As a patient nears discharge, the model could be leveraged in making the decision to discharge and arranging post-discharge support. Hospitals that follow-up with patients during the transition of care between discharge and return to the medical home could leverage risk models to determine how long to monitor patients. Home health services that send nurses to patients' homes could use risk models to understand and identify high-risk patients among those they follow. We believe that similar risk models also might be applicable to other adverse events such as complications of surgery and response to therapy.

We present a methodology for developing predictive models for hospital readmission within 30 days that incorporate semantically meaningful derived data elements representing phenotypes. We leverage an existing software system, the Analytic Information Warehouse (AIW)[3, 4], which computes derived variables from clinical data that are specified as groups of codes, or frequency, sequential and other temporal patterns. This system has access to five years of clinical and administrative data from our institution's clinical data warehouse. We capture socio economic information through geocoding of addresses and use of U. S. Census (http://www.census.gov/) and American Community Survey (http://www.census.gov/acs/www/) data. We have carried out logistic regression to explore the strength of association of individual and combinations of derived variables with 30-day readmissions in different patient populations. We have applied the random forest algorithm[5] to develop predictive models for these populations. These models are intended for ultimate use as components of clinical decision support post-discharge and upon admission.

Background

Major risk factors that have been found to be associated with hospital readmission within 30 days include age, race, sex, co-morbidities, economic disadvantage, drug use, income, and number of previous hospitalizations[6, 7]. These data elements are available in EHRs to varying degrees. While socioeconomic variables are not widely available in structured form in EHRs as described above, aggregate variables from census data can be linked to patient records through the patient's address. Co-morbidities are primarily available via billing diagnosis codes. These may only be recorded by coders after discharge and thus may not be available for prediction during a hospitalization except for codes recorded in previous encounters. Diagnoses and conditions likely can be inferred in some cases from medication history, laboratory test results and/or procedures. EHR data may be quite useful in readmissions prediction for causes of readmissions that manifest themselves in information that can be extracted or inferred from EHRs and clinical data warehouses.

Readmission predictive models using subsets of these variables have been built to facilitate calculation of risk-standardized readmission rates for comparing hospital quality of care. They also have been implemented to compute risk for specific patient populations with the goal of reducing readmission rates and costs[2, 6]. Kansagara et al.[2] conducted a systematic review on risk prediction models for hospital readmission and revealed that there were only 3 models, developed and tested in large European or Australian cohorts, with area under the receiver operating characteristic curve (AUC) of 0.70 or higher. AUC of 0.5 indicates that performance of the model is no better than chance. For the US-based studies, 9 models showed AUC ranging from 0.55 to 0.65, indicating that the discriminative ability was poor. Models have tended to perform worse outside of the institution at which they were developed. Possible explanations include the multifactorial nature of potential causes of readmissions, the challenges of using EHR data for this purpose as described earlier, attempting creation of a single general-purpose model rather than building different models for different broad classes of patients, and/or not leveraging socioeconomic data.

Most predictive models[2] have been designed to make a prediction after the patient has been discharged. Amarasingham et al.[6] developed a model to identify heart failure patients at high risk of readmission or death using pre-discharge data in order to provide guidance to clinicians while a patient is still in the hospital. Their model performed better than the heart failure readmission and mortality models developed by the Center for Medicaid and Medicare Services (CMS models)[8] and a mortality model derived from the Acute Decompensated Heart Failure Registry (ADHERE model)[9]. Still, performance of these models has been limited.

National projects have embarked upon efforts to create reusable clinical phenotypes that are specified in terms of EHR data. Computing such phenotypes has emerged as a requirement for systematic study of the relationships between disease, progression and therapeutic response and genotype in gene-based disease management. One such project, eMERGE[10], is making its phenotype definitions publicly available for reuse, though as documents rather than in computable form. Other projects, such as the AIW (see above), are constructing tools for specifying derived variables such as phenotypes in a reusable form that allows for differences in data representation in various EHRs and clinical data warehouses. Repositories of phenotypes combined with tools for representing them in computable form in a broadly applicable way together are expected to enable their use in a broad array of investigations including the development of predictive models. We have specified using the AIW over 50 derived variables representing different clinical phenotypes of interest in our readmissions prediction efforts. The details of our methods are described below.

Methods

The development and evaluation of models for characterizing and predicting readmissions within 30 days has included the following data retrieval, derived variable computation, and model development steps. While this work was performed as part of hospital operations, we obtained Institutional Review Board approval to generalize and publish it.

Data

We extracted all hospital encounters between 2006 and 2011 from our institution's clinical data warehouse for modeling. We retrieved demographics, encounter (as defined by billing systems) location, encounter start and end dates, discharge status codes (a representation of post-discharge care), ICD-9 diagnosis and procedure codes, MS-DRG codes, and selected laboratory test results, medication orders and vital signs. The extract contained 238,996 encounters in 149,514 patients. We used AIW software (described below) to compute derived variables representing phenotypes specified as groups of codes, thresholds in test results and sequential and frequency temporal patterns (selected phenotypes are shown in Table 2). We removed all encounters that did not have a billing discharge diagnosis code in at least one of 10 disease categories (shown in left-most column of Table 1) that are associated with elevated risk of hospital readmission within 30 days. We also excluded encounters from analysis that were related to childbirth or if the subsequent hospital encounter was a rehabilitation, chemotherapy, radiation therapy, or psychiatry encounter (selected by ICD-9 codes or the location of the encounter). The latter exclusion criteria aimed to remove anticipated or unpreventable encounters from analysis. Using this process, we extracted 10 datasets, one for each of the 10 disease categories. Numbers of encounters in each dataset are shown in Table 1.

Table 1. Number of encounters available for creating predictive models for ten disease categories (see Table 2 for selected definitions). Number of encounters in patients with at least 2 encounters is shown because for training models for predicting risk at the start of a hospital encounter ("on admission", see Methods), we withheld all data from the "current" encounter, thus requiring at least two encounters to be present.

Disease Category	Number of encounters	Number of encounters in patients with at least 2 encounters
Cancer	50,391	19,254
Chronic kidney disease	41,901	21,843
Chronic obstructive pulmonary disease	21,203	8,650
Diabetes	55,093	22,730
Heart failure	21,550	17,461
Myocardial infarction	22,403	7,645
Pulmonary Hypertension	5,973	2,897
Sickle cell anemia	2,186	1,561
Stroke	6,858	1,705
Transplant	5,147	2,690

Table 2. Selected derived variables used in readmissions analyses.

Variable name	Definition	Variable name	Definition
30-day readmission	A hospital encounter with an admit date within 30 days of the previous hospital encounter's discharge date	Chronic kidney disease (CKD)	At least one of the following billing ICD-9 codes, either primary or secondary: 581.*, 582.*, 585.*
Sickle cell crisis	At least one of the following billing ICD-9 codes, either primary or secondary: 282.62, 282.64	Sickle cell anemia	At least one of the following billing ICD-9 codes, either primary or secondary: 282.6*
End-stage renal disease (ESRD)	At least one of the following billing ICD-9 codes, either primary or secondary: 285.21, 585.6	Myocardial infarction	At least one of the following billing ICD-9 codes, either primary or secondary: 410.*
Chemotherapy 180 days before surgery	Chemotherapy encounter discharge date at most 180 days before a surgical procedure	Radiation therapy encounter	Primary or secondary billing diagnosis code V58.0
Chemotherapy encounter	At least one of the following primary or secondary billing diagnosis codes: V58.1*	Multiple MIs	Two or more *Myocardial infarction*s across all encounters for a patient.
Pressure ulcer	1) At least one of the following ICD-9 codes, either primary or secondary: 707.0, 707.2	Cancer	At least one of the following billing ICD-9 codes, either primary or secondary: 140-208, 209.0, 209.1, 209.2, 209.3, 225.*, 227.3, 227.4, 227.9, 228.02, 228.1, 230.*, 231.*, 232.*, 233.*, 234.*, 236.0, 238.4, 238.6, 238.7, 239.6, 239.7, 259.2, 259.8, 273.2, 273.3, 285.22, 288.3, 289.83, 289.89, 511.81, 789.51, 795.06, 795.16, V58.0, V58.1*, V10.*
Metastasis	At least one of the following billing ICD-9 codes, either primary or secondary: 196.*, 197.*, 198.*	Heart failure from BNP	B-type natriuretic peptide (BNP) test result 100-300 = suggest heart failure is present; 300-600 = mild heart failure; 600-900 = moderate heart failure; or > 900 = severe heart failure
Surgical procedure	ICD-9 procedure codes 01-86.99		

Derived Variable Computation

The AIW software is based on a previous software system, PROTEMPA[11], which uses the temporal abstraction method[11, 12] to compute temporal patterns and classifications in clinical data. The AIW leverages PROTEMPA's capabilities for specifying and computing derived variables as categories of codes, classifications of continuous variable values, and frequency, sequential and other temporal patterns in time-stamped data. Derived phenotypes are specified in a temporal abstraction ontology[3, 12] using the Protégé editor and its frame-based ontology format (http://protege.stanford.edu). The AIW supports connecting to a clinical data warehouse, generating and executing SQL queries for data needed to determine when a patient is expressing a specified phenotype, scanning for categories, classes and patterns, and loading data and derived variables into existing analysis tools. The AIW was configured for use in this study with phenotype specifications for the disease categories of interest (Table 1), a specification of a readmission within 30 days for any cause (Table 2), and other co-morbidities and disease conditions (Table 1 and Table 2). The AIW provided output as a delimited file with one hospital encounter per row, and several hundred encounter attributes as columns. The ten datasets described above were extracted from AIW output (see *Data*).

Logistic Regression

We used logistic regression computed with SAS 9.2 (SAS Institute Inc., Cary, NC) to create an initial characterization of relationships between readmission rate and combinations of derived variables. Odds ratios and confidence levels were computed to reveal factors having a substantial impact on the readmission rate. We then went on to develop predictive models using these factors and random forest algorithm[5], described below, which is known to handle missing values and scale well to developing models using large multivariate datasets. The random forest algorithm can identify additional variables that are associated with readmissions beyond those found through regression analysis.

Random Forest

The random forest algorithm generates a predictive model by constructing an ensemble of decision trees[13] based on the training data. A decision tree is learned by partitioning the training data based on the values of the variable that leads to most information gain. This process of partitioning is applied recursively using the rest of the variables to generate a set of rules arranged in tree form with partial overlap of *if-then* rules. The random forest algorithm constructs each tree using a randomly drawn subset of the variables in the data and using a subset of examples from the training set. This creates a set of trees each with a slightly different model. The combination of this ensemble leads to a robust predictive model. To classify a hospital encounter's risk, an election is conducted where each tree casts a vote *yes* or *no* for whether a subsequent readmission within 30 days will occur, and a risk value is computed as the fraction of trees that voted *yes*. Initial experiments strongly suggested that disease-specific models perform better than a single model built with all data. Thus, we developed separate models for the ten disease categories described above (Table 1). We developed two models for each category representing two of the use cases described above – providing decision support post-discharge and on admission. For the post-discharge models, we included diagnosis and administrative codes, lab values and medications that are available through discharge. For on admission models, we only considered those data from prior encounters.

Random forest calculations were performed using freely available WEKA data mining software[14] (http://www.cs.waikato.ac.nz/ml/weka/) and R (http://www.r-project.org/). We conducted 5-fold cross validation for each model. In each fold, random forest used 4 sets to generate the ensemble of trees and tested on the remaining set. The trained random forests generated a probability of readmission within 30 days for any cause for each encounter in the test set, and the encounters were ranked from high risk to low risk.

Results

Logistic Regression Results

Over 50 pairs of derived variable values had statistically significant ($p < 0.05$) increased risk of readmission. For cancer patients, those with multiple heart attacks were approximately 2.2 times more likely to be readmitted to the hospital within 30 days than those who did not have multiple heart attacks after controlling for other variables. Also for cancer patients, those with metastasis were approximately 1.3 times more likely to be readmitted to the hospital within 30 days than those without metastasis after controlling for other variables. For chronic kidney disease patients, those with End Stage Renal Disease (ESRD) were approximately 1.4 times more likely to be readmitted to the hospital within 30 days than those without ESRD after controlling for other variables. Sickle cell patients living in a county with a low average house value as recorded in the American Community Survey were at higher risk of being readmitted to the hospital within 30 days than sickle cell patients living in a county with a high average house value. Selected statistically significant pairs are reported in Table 3.

Table 3. Logistic Regression Analysis of Selected Derived Variables. See Table 2 for definitions of disease categories and derived variables.

Disease category	Derived variable	Level	Odds Ratio	*p* value
Cancer	Metastasis	Yes vs. No	1.331	0.001
	Multiple MIs	Yes vs. No	2.188	< 0.0001
Transplant	Pressure Ulcer	Yes vs. No	3.324	0.004
Myocardial Infarction	Heart Failure from BNP	Suggest_heart_failure_is_present vs. No_heart_failure	1.096	0.0289
		Indicate_mild_heart_failure vs. No_heart_failure	1.205	
		Indicate_moderate_heart_failure vs. No_heart_failure	1.295	
		Indicate_severe_heart_failure vs. No_heart_failure	1.422	
Sickle Cell Anemia	County Average House Value	<= $87,400 vs > $140,000	1.594	< 0.0001
		$87,400 - $101,800 vs > $140,000	2.153	
		$101,800 - $140,000 vs > $140,000	1.867	
Chronic Kidney Disease	End-stage Renal Disease	End-stage Renal Disease vs. Chronic Kidney Disease	1.395	< 0.0001

Random Forest Results

Random forest predictive model generates a ranking of a given set of patients from highest risk for 30-day readmission to lowest risk. Results for the post-discharge and on admission models are shown in Table 4 and Table 5, respectively. Results shown are the average of the 5 folds. For the 10 disease categories, 5 fold test set average ("baseline") of the 30-day readmission rate ranged from 12% to 34%. For each of the 10 disease categories, we show the relative difference in readmission rates for the highest (top decile) and lowest (bottom decile) risk groups as compared with the average rate ("baseline") for that disease category expressed as ratios. For every disease category in the on admission and post-discharge scenarios, the readmission rate of the highest risk decile was substantially greater than the baseline readmission rate for that category. Similarly, for all categories in both scenarios, the readmission rate of the lowest risk decile had a lower readmission rate than its corresponding baseline rate.

Table 4. Post-discharge analysis: relative differences in all-cause rates of hospital readmission within 30 days for predictive modeling high and low risk groups as compared with baseline rate for each of 10 disease categories.

Diagnosis Category	Ratio of Readmission Rate in Predictive Modeling Group versus Baseline Readmission Rate	
	Predictive Modeling High Risk Group	Predictive Modeling Low Risk Group
Cancer	2.6610	0.2392
Chronic kidney disease	2.2941	0.2448
Chronic obstructive pulmonary disease	2.5874	0.2934
Diabetes	2.6659	0.2337
Heart failure	2.4869	0.2298
Myocardial infarction	2.9349	0.1147
Pulmonary hypertension	2.2256	0.1459
Sickle cell anemia	2.0678	0.2496
Stroke	3.9265	0.1667
Transplant	2.1347	0.4014

Table 5: On admission analysis: relative differences in all-cause rates of hospital readmission within 30 days for predictive modeling high and low risk groups as compared with baseline rate for each of 10 disease categories.

Diagnosis Category	Ratio of Readmission Rate in Predictive Modeling Group versus Baseline Readmission Rate	
	Predictive Modeling High Risk Group	Predictive Modeling Low Risk Group
Cancer	1.9783	0.3018
Chronic kidney disease	1.8473	0.3744
Chronic obstructive pulmonary disease	2.0979	0.4911
Diabetes	2.0186	0.3656
Heart failure	2.0123	0.3775
Myocardial infarction	2.1953	0.3468
Pulmonary hypertension	1.8315	0.4410
Sickle cell anemia	1.5548	0.2376
Stroke	2.4180	0.3589
Transplant	1.7573	0.4610

Discussion

We present results of associative and predictive modeling for understanding clinical phenotypes that are expressed in hospitalized patients who are readmitted within 30 days and identifying patients who are at risk for a readmission.

The associative studies are intended for quality improvement analysts and investigators to obtain an enhanced understanding of the causes of readmissions. The predictive models that are developed based on insights derived from found associations are aimed ultimately at providing clinical decision support at the point of care.

These models were developed and tested using retrospective EHR and census tract data. Logistic regression analysis in combination with high-level phenotypic patient descriptions (Table 3) yielded easily understandable comparisons of the strength of association of various phenotypes with all-cause readmission within 30 days. Identifying which phenotypes appear correlated with our outcome variable is an important step in selecting phenotypes for inclusion in predictive models. For predictive model construction, we were able to use a well-understood machine learning algorithm, random forest, while retaining the ability to incorporate temporal and categorical information in our analyses through the use of derived variables.

For predictive model evaluation, we simulated model use either upon admission (Table 5) or after discharge (Table 4). In both scenarios, our approach was successful, though not perfect, in identifying high risk and low risk groups of patients. A key part of this success was limiting our predictions to those encounters in the top and bottom risk deciles. Intuitively, because the models have access only to incomplete information as described above, it seems appropriate to limit our predictions to those cases in which the EHR has relatively clear evidence of risk. Developing separate models for different broad disease categories (Table 1) performed better than a single all-purpose model, likely reflecting that the etiologies of readmissions that are accessible in EHRs vary by disease.

The derived variable phenotypes were specified and computed using a software system, the AIW, that has been described previously for use in computing derived variables in comparative research studies. These phenotypes may be specified as groups of codes or temporal patterns. We expect that projects that are defining phenotypes for use in research (see above) are synergistic with similar efforts such as this to define phenotypes of value in quality improvement studies. Both require the same mechanism for specifying phenotypes in computable form and identifying patients who express them. The AIW software has been valuable in computing derived variables in both domains. Both also would benefit from being able to specify phenotypes in a computable form that allows reuse across institutions. Reusable phenotype specifications potentially could enable sharing of predictive models as well. The requirements for achieving such reuse have yet to be determined.

The derived variables in this work (Table 2) were chosen in collaboration with Emory clinicians, including a group of nurse practitioners called “transition managers” who are responsible for tracking patients at risk for readmission while in the hospital and post-discharge until they receive follow-up care from their medical home. These variables have defined sensitivity and specificity profiles that are important to know in interpreting our analyses and predictions. In specifying variables, we prioritized high sensitivity under the assumption that clinicians using these models would have additional information that could quickly exclude a patient from further consideration. Thus, we did not apply common constraints on such phenotypes such as requiring more than one of a diagnosis code or multiple sources of evidence for a condition. Work is needed to refine these phenotypes and analyze their accuracy.

In both the on admission and post-discharge use cases, we believe that these models could produce additional information for the patient care team in a clinical decision support capability that is intended to target limited resources to those patients who need it most. As health systems increasingly track patient status post-discharge, we expect clinical decision support for the period between hospitalization and the first follow-up appointment to grow in importance. Deploying our models in the hospital and post-discharge environments is an area of future work. We anticipate being able to optimize our models further by improvement of our algorithms, enhancement of our existing derived variables and incorporation of new ones. Increasing our use of clinical data to infer the patient’s disease state such as through medication history or test results is an area of future work that we expect to enhance the accuracy of our models especially in the on admission use case where current billing diagnosis codes may be unavailable. With these optimizations in place, substantial effort would be required to deploy and maintain predictive models in existing EHRs as a form of clinical decision support that presents clinicians with a patient-specific risk at the point of care.

Conclusion

Computational models of hospital readmissions within 30-days can provide valuable information in identifying variables associated with readmissions and identifying patients at high and low risk. Computing derived variables that represent clinical phenotypes and leveraging them in risk models may enhance models' accuracy and make them reusable in a wider variety of clinical environments. Clinical decision support tools that leverage these models in providing risk assessments at the point of care could enhance decision-making during hospitalization and during the post-discharge period.

References

1. QualityNet. Readmission Measures Overview: Publicly reporting risk-standardized, 30-day readmission measures for AMI, HF and PN. Accessed Mar 1, 2012. http://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1219069855273
2. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA : the journal of the American Medical Association. 2011;306(15):1688-98. Epub 2011/10/20.
3. Post A, Kurc T, Overcash M, Cantrell D, Morris T, Eckerson K, et al. A Temporal Abstraction-based Extract, Transform and Load Process for Creating Registry Databases for Research. AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science. 2011;2011:46-50. Epub 2012/01/03.
4. Post AR, Sovarel AN, Harrison JH. Abstraction-based temporal data retrieval for a Clinical Data Repository. AMIA Annu Symp Proc. 2007:603-7.
5. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
6. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Medical care. 2010;48(11):981-8. Epub 2010/10/14.
7. Kim H, Ross JS, Melkus GD, Zhao Z, Boockvar K. Scheduled and unscheduled hospital readmissions among patients with diabetes. The American journal of managed care. 2010;16(10):760-7. Epub 2010/10/23.
8. Keenan PS, Normand SL, Lin Z, Drye EE, Bhat KR, Ross JS, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation Cardiovascular quality and outcomes. 2008;1(1):29-37. Epub 2008/09/01.
9. Fonarow GC, Adams KF, Jr., Abraham WT, Yancy CW, Boscardin WJ. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. JAMA : the journal of the American Medical Association. 2005;293(5):572-80. Epub 2005/02/03.
10. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011;4:13. Epub 2011/01/29.
11. Post AR, Harrison JH, Jr. PROTEMPA: A Method for Specifying and Identifying Temporal Sequences in Retrospective Data for Patient Selection. J Am Med Inform Assoc. 2007;14:674-83.
12. Shahar Y. A framework for knowledge-based temporal abstraction. Artif Intell. 1997;90:79-133.
13. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, California, USA: Wadsworth, Inc; 1984.
14. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10-8.