A Machine Learning Approach to College Drinking Prediction and Risk Factor Identification

JINBO BI, JIANGWEN SUN, and YU WU, University of Connecticut
HOWARD TENNEN, University of Connecticut Health Center
STEPHEN ARMELI, Fairleigh Dickinson University

Alcohol misuse is one of the most serious public health problems facing adolescents and young adults in the United States. National statistics shows that nearly 90% of alcohol consumed by youth under 21 years of age involves binge drinking and 44% of college students engage in high-risk drinking activities. Conventional alcohol intervention programs, which aim at installing either an alcohol reduction norm or prohibition against underage drinking, have yielded little progress in controlling college binge drinking over the years. Existing alcohol studies are deductive where data are collected to investigate a psychological/behavioral hypothesis, and statistical analysis is applied to the data to confirm the hypothesis. Due to this confirmatory manner of analysis, the resulting statistical models are cohort-specific and typically fail to replicate on a different sample. This article presents two machine learning approaches for a secondary analysis of longitudinal data collected in college alcohol studies sponsored by the National Institute on Alcohol Abuse and Alcoholism. Our approach aims to discover knowledge, from multiwave cohort-sequential daily data, which may or may not align with the original hypothesis but quantifies predictive models with higher likelihood to generalize to new samples. We first propose a so-called temporally-correlated support vector machine to construct a classifier as a function of daily moods, stress, and drinking expectancies to distinguish days with nighttime binge drinking from days without for individual students. We then propose a combination of cluster analysis and feature selection, where cluster analysis is used to identify drinking patterns based on averaged daily drinking behavior and feature selection is used to identify risk factors associated with each pattern. We evaluate our methods on two cohorts of 530 total college students recruited during the Spring and Fall semesters, respectively. Cross validation on these two cohorts and further on 100 random partitions of the total students demonstrate that our methods improve the model generalizability in comparison with traditional multilevel logistic regression. The discovered risk factors and the interaction of these factors delineated in our models can set a potential basis and offer insights to a new design of more effective college alcohol interventions.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Feature evaluation and selection*

General Terms: Algorithms, Design, Performance

Additional Key Words and Phrases: Machine learning, longitudinal data analysis, feature selection, clustering, college student alcohol consumption

An early result of our analysis was published in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium.*

Authors' addresses: J. Bi, Computer Science and Engineering Department, University of Connecticut, and Connecticut Institute for Clinical and Translational Science; email: jinbo@engr.uconn.edu; J. Sun and Y. Wu, Computer Science and Engineering Department, University of Connecticut; H. Tennen, Community Medicine and Healthcare Department, University of Connecticut Health Center; S. Armeli, Psychology Department, Fairleigh Dickinson University.

DOI: http://dx.doi.org/10.1145/2508037.2508053

ACM Reference Format:
Bi, J., Sun, J., Wu, Y., Tennen, H., and Armeli, S. 2013. A machine learning approach to college drinking prediction and risk factor identification. ACM Trans. Intell. Syst. Technol. 4, 4, Article 72 (September 2013), 24 pages.
DOI: http://dx.doi.org/10.1145/2508037.2508053

1. INTRODUCTION

Alcohol use among adolescents and young adults is an important public health issue [Chung et al. 2005; Maggs and Schulenberg 2005; Wechsler et al. 1995b]. Existing research indicates that roughly 80% of college students drink alcohol and that at least half of college student drinkers engage in heavy episodic drinking [Carter et al. 2010]. Studies have supported that college students who participate in excessive alcohol intake are more likely to become involved in adverse activities resulting in negative consequences [Simons et al. 2005; National Institute of Alcohol Abuse and Alcoholism 2002; Hingson et al. 2009, 2002]. These consequences range from nonfatal or fatal injuries, alcohol poisoning, blackouts, academic failure, violence (which includes rape and assault), and sexually transmitted diseases, to criminal activities that may jeopardize future life prospects [National Institute of Alcohol Abuse and Alcoholism 2002; Hingson et al. 2009, 2002; Toumbourou et al. 2009]. The adverse outcomes affect not only the students themselves but also fellow students, roommates, and family members, in some cases, as the secondary effect [Moulton et al. 2000; Room 1996; Toumbourou et al. 2009].

Binge drinking is one of the most serious college drinking problems [National Institute on Drug Abuse 1999]. It is estimated that roughly 90% of the alcohol consumed by youth under 21 years of age in the United States is in the form of binge drinks [Office of Juvenile Justice and Delinquency Prevention 2005]. College binge drinkers are more likely than nonbingers to exhibit a wide range of problem behaviors, including illicite drug use and drunk driving [Wechsler et al. 1995a]. Binge drinking is defined as episodic excessive drinking [Stolle et al. 2009]. Currently there is no international consensus on how many drinks constitute a *binge* [International Center for Alcohol Policies 2003]. In the United States, the term “binge” often refers to as consuming five or more standard drinks per occasion for men and four or more drinks per occasion for women [Wechsler and Austin 1998; Wechsler and Nelson 2001].

Alcohol addiction and dependence is a medical problem characterized as a chronic, often progressive disease with symptoms that include a strong need to drink despite negative consequences [American Psychiatric Association 1994]. Like many other diseases, it has a generally predictable course, recognized symptoms, and is influenced by factors both genetic and environmental [Jellinek 1960]. As college students age beyond college, many exhibit a tendency to moderate heavy alcohol involvement [Perkins 1999; Donovan et al. 1983]. However, against this normative decrease, some young adults continue to drink heavily and manifest alcohol use disorders [Zucker et al. 1995]. Thus a central question of research into the etiology of alcoholism concerns the identification of factors that distinguish young adults who moderate their alcohol consumption from those who persist in heavy drinking [Jackson and Sher 2005]. Numerous studies have focused on the recognition of risk and resilient factors for alcohol-related problems. The factors hypothesized as predictors of college drinking comprise, in many separate studies, the motivational pathways [Stewart et al. 2006; O’onnor and Colder 2005], developmental pathways [Auerbach and Collins 2006], epidemiologic distributions [O’Malley and Johnston 2002], and other theorized predictors of alcohol use. The data observed in a study that is designed to prove a specific hypothesis; however, may carry potential information and utility beyond the original hypothesis [National Institutes of Health 2008]. More thorough and accurate analysis of the data has the potential to expand the scientific understanding of alcohol use behaviors.

Longitudinal studies have emerged to investigate college student high-risk drinking and its longitudinal course [Jackson and Sher 2005; Timberlake et al. 2007; Armeli et al. 2010; Stappenbeck and Fromme 2010; Turner et al. 2010; Walls et al. 2009]. Through these studies, a number of risk factors for alcohol misuse have been identified that may be relevant to understanding the onset and course of binge drinking. Especially, daily study designs capture temporal dynamics of rapidly fluctuating processes, such as mood and coping, close to their real-time occurrence [Tennen et al. 2000; Armeli et al. 2003, 2010; DeHart et al. 2009]. They offer unique insights into challenging questions of college drinking. However, although the daily diary itself may capture the daily dynamics of a process, the statistical methods used to analyze the data seldom examine the dynamics at the daily level. Instead, traditional methods, such as multilevel regression or multilevel logistic regression, model the data by averaging daily measures to account for an overall daily effect that is then nested at the month, year, person level [Singer and Willett 2003]. As the resulting models are only informative about daily averages of drinking, moods and stress, they do not address within-person association or contingency [Armeli et al. 2000].

We perform a secondary analysis on data of a recently-completed survey collected in a college drinking motive study at University of Connecticut Alcohol Research Center sponsored by National Institute of Alcohol Abuse and Alcoholism. Two sets of survey data were gathered: a baseline survey was used to collect a student's personality traits, motives and other person-level variables, and a daily survey utilized in a period of continuous 30 days was used for students to report daily stress, moods, emotions, drug use behavior, and social behavior. Specific objectives of our secondary analysis are to provide support for (1) analyzing previously collected data that would advance, in cost-effective ways, scientific knowledge of problem drinking behavior, and (2) applying new approaches to analyze current datasets that identify patterns and hypotheses to benefit from further exploration.

In our analysis, two categories of machine learning algorithms are studied. Supervised classification algorithms construct mathematical models that separate samples into different categories, where categories are predetermined by human experts and category labels are pre-assigned to training samples. Unsupervised cluster analysis develops classification labels automatically as part of the algorithm process. This type of algorithm has to search for similarities between subsets of data in order to determine whether or not they can be characterized as forming a group. For supervised learning, training performance of a model, defines the accuracy that the model predicts the categories of the data that is used to train the model. Nevertheless, generalization performance measures the accuracy that the model predicts on data that is independent of the training data. The classic cohort-specific criteria used in alcohol literature correspond to only the training performance of a model, as all collected data has been used to build the model. Our secondary analysis of the drinking-related student diaries contributes uniquely as follows.

(1) For longitudinal data analysis, an algorithm dealing with repeated measurements is devised based on support vector machines and employed in the analysis of the daily diaries of stress, moods, and other daily variables. The algorithm aims to construct models to classify days when a student had nighttime binge drink episodes from days when s/he did not. The models, once validated, can be used to predict the vulnerability of any binge drinking occurring on a given night based on daily measures. Our algorithm models the daily dynamics directly instead of averaged daily effects. It employs a sparse formulation which targets a model that uses the least

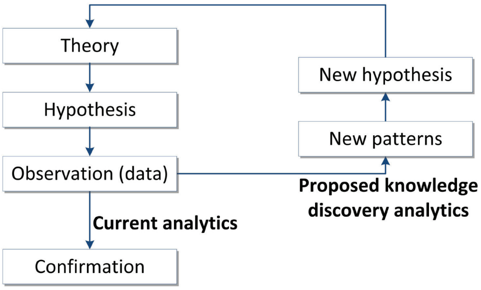

Fig. 1. Deductive versus inductive analytics. Current analytic methods are used to confirm known hypothesis (deductive), and the proposed analytic methods aim to discover unknown but potentially insightful hypothesis (inductive).

number of daily variables to best predict the occurrence of nighttime binge drinking. The models quantify the interplay of the selected variables, thus emphasizing the *interactive effects* of the selected factors on drinking.

(2) For recognition of alcohol use patterns, we employ a consecutive combination of cluster analysis followed by statistical feature selection. In cluster analysis, we search for partitions of the study subjects based on their drinking variables, which then forms the severity categories of alcohol use. Feature selection is used to identify designated personality factors highly discriminative between categories in the partition. The methodology of incorporating an unsupervised cluster analysis facilitates to discover new behavior patterns without the bias of the original hypothesis for which the data was collected.

(3) The evaluation of both these methods focuses on the generalization performance of the resulting models on test samples. The proposed methods are compared against the traditional approach of logistic regression with respect to the generalization performance. A higher generalization performance of a model implies higher likelihood for the model to be replicated on distinct cohorts. The daily variables identified by the first methodology will help explain the short-term daily process of binge drinking with proximal contingency on previous three days. The risk factors identified by the second methodology will constitute the person-level trait characteristics, long-term coping strategies and drinking motives, which may be relevant to understanding the diversing path between students at their young adult years in regards of the capability of moderating alcohol involvement. These factors may bring insights into the etiology of alcoholism and benefit the design of new interventions.

The rest of the article is organized as follows. Section 2 provides the background and related literature of our problem and analysis. Section 3 describes the specifics of the data used in our analysis. We discuss the proposed machine learning algorithms and their results in Section 4. Section 4 comprises two sections, one for the temporally-correlated support vector machine that analyzes repeated daily measurements, and the other for composite analytics that models the person-level baseline variables. Experimental results and interpretations are provided in the respective subsections. Section 5 discusses the implication of our findings for future research and concludes.

2. BACKGROUND AND RELATED WORK

Statistical analytics is widely used in the alcohol literature to test research hypothesis. The standard paradigm of data analysis, as shown in Figure 1 (left half), is a deductive process that derives hypothesis from theory, and a study is designed to prove the hypothesis by collecting data. Statistical methods are used to analyze the data for confirmation of the original hypothesis. Current alcohol studies are hampered by the confirmatory nature of their analytic methods. Our contribution, as discussed in the previous section and shown in Figure 1 (right half), includes the design of machine-learning algorithms that aim to discover new patterns without the bias of original hypotheses, potentially leading to new exploratory hypotheses which may be cross-validated on existing data or in a new study. The design of our algorithms may be guided by psychological theories, but the discovery is data-driven by applying the proposed algorithms to an existing set of alcohol data.

2.1. Existing Analytic Methods for Risk Factor Identification

The analytic techniques used to validate hypothesized risk factors range from descriptive statistics (e.g., mean and standard deviation) and parametric inference models to factor analysis. Parametric inference models, such as logistic regression [Beck et al. 2008], linear regression or hierarchical linear regression [Singleton and Wolfson 2009; DeJong et al. 2009], have been commonly used. Negative binomial regression [Lewis et al. 2009], multinomial logistic regression [Reed et al. 2007], and multilevel regression [Patrick and Lee 2010] are among the more elaborated methods when sample data presents complex structure. Factor analytic methods [Talbott et al. 2010; LaBrie et al. 2010] can potentially identify interactions among correlates or among the dependent variables. They are also frequently employed in conjunction with inference models in hypothesis testing. More comprehensive methods, such as structural equation modeling [Abar and Maggs 2010], embrace factor analysis and regression as special cases.

In all of the cited studies, no research data was used separately to test the models that were derived from the recruited sample. For example, Singleton and Wolfson [2009] conducted interview surveys with a random sample of 236 students to examine the associations between sleep, alcohol use, and academic performance in college students. The interviews measured alcohol consumption, gender, academic class, weekday and weekend bedtimes and rise times, which similarly, our diary data also measured. Multiple linear regressions were applied to the entire dataset and reported p-values that showed alcohol consumption was a significant predictor of specific sleep patterns, and it had direct and indirect relationships with the grade point average (GPA). Although linear models were built with multiple independent variables, the p-value was reported only for each individual variable, and the variables with small p-values were considered as predictors. There exist two drawbacks in this approach. The analysis did not specify the interplay between the identified predictors, although the linear models themselves could quantify it. These models were not tested on any independent sample to see if the models could generalize to its study population.

2.2. Existing Analytic Methods for Longitudinal Studies

Analysis of repeated measurements obtained in longitudinal studies plays an increasingly prominent role in college drinking research to reveal the within-person dynamics of alcohol use and how these dynamics vary among individuals. For longitudinal data analysis, the most widely used statistical methods are a family of conditional and unconditional multilevel models [Singer and Willett 2003; Armeli et al. 2010], including multilevel logistic regression and multilevel linear regression in which, for instance, drinking days are nested within months, which are then nested within persons. Path and trajectory analysis [Sher et al. 2011] has also been used to model longitudinal data to identify patterns of behavioral trends and transitions over the study period. Transitional linear models, such as latent transition analysis [Auerbach and Collins 2006] and latent curve analysis [Wood et al. 2010], treat alcohol use as a dynamic latent variable and categories, such as “non-drinker”, or “heavy drinker”, as latent classes of the latent variable, and estimates the probabilities of the latent classes at each occasion together with the transition probabilities.

We argue two major limitations of current approaches. First, similar to the issues in Section 2.1, the statistical analysis has focused on modeling data that has been collected instead of forecast prediction over time. Whereas a modeling process takes as its goal determining how accurately the model fits the collected data, a predictive inference takes as its goal determining how likely the resulting model can predict future cases. The terminology "overfitting" in statistical learning theory elucidates the difference of the two goals [Vapnik 1995]. Since random variability always exists in study samples, especially in sample sizes commonly used in alcohol use studies, an inference model that matches the observed data perfectly does not necessarily correspond to the underlying mapping from covariates to the dependent variable as the model fits the random noise as well. Second, many study designs collect data to catch rapidly fluctuating processes, but their analysis does not utilize the fluctuation dynamics, such as in the prior analysis of our daily diaries where multilevel models investigated average daily effects rather than day-to-day dynamics or transitions. Section 2.3 provides more details on the prior study of our data.

2.3. Prior Analysis of Our Study Data

The dataset used in our analysis was previously analyzed using multilevel regression methods for investigating if drinking motives moderated the association between daily negative affect and alcohol use among college students [Armeli et al. 2010]. Although many perspectives of student daily life were measured, resulting in 56 daily variables repeatedly measured for 30 days, only five variables related to drinking and daily negative affect were used in the analysis. Two measures "sad" and "dejected" were used to create a variable for *daily depressive affect*. Two measures "jittery" and "nervous" were used to create a variable for *daily anxious affect*. Another daily variable, "the amount of drinks each day", was used to create two variables, the *drinking frequency* and the *average amount of daily drinks*. Besides the five daily variables, sex, age, and six person-level measures, *drinking to cope*, *drinking to socialize and enhance*, *Beck depression inventory index*, *Spielberger trait anxiety inventory*, and *retrospective drinking frequency* and *intensity* in the past 30 days were utilized. Daily variables were averaged over the 30 days in the study period. The negative affect variables were centered by the person-level means. A three-level logistic model was constructed using the daily diaries with days nested within years, which were nested within persons. The models at different levels were neither tested on a second cohort nor tested in a cross-validation procedure.

Logistic regression is a widely used method for binary classification problem, where each subject receives a label of either $y = 1$ or $y = -1$ [Hastie et al. 2001]. Its model is given by $p(y|\mathbf{w}, \mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+exp(-y\mathbf{w}^{\mathbf{T}}\mathbf{x})}$, where $\mathbf{x}$ denotes the features describing a subject and $\mathbf{w}$ is the model parameter to be determined. The likelihood of observing n samples of $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, can be calculated as $L(\mathbf{y}|\mathbf{w}, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \Pi_{i=1}^{n} \frac{1}{1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)}$. Regular logistic regression maximizes the likelihood L for the optimal $\mathbf{w}$, which correponds to minimizing the negative logarithm of the likelihood. The posterior distribution of the model parameter $\mathbf{w}$ is proportional to the product of the likelihood and a prior distribution $p(\mathbf{w})$, that is, $p(\mathbf{w}|\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \Pi_{i=1}^{n} \frac{1}{1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i)} p(\mathbf{w})$, which is used and maximized in the maximum-a-posterior version of logistic regression. For numerical convenience, sometimes a tuning parameter c is used to balance between the prior and the likelihood during determination of the optimal $\mathbf{w}$.

Table I. Person-Level Measures/Factors

Factor ID	Question/Factor
1	Drinking to Cope
2	Drinking to Socialize
3	Drinking to Enhance
4	Drinking to Conform
5	Sensation Seeking
6	Neuroticism
7	Beck Depression Inventory
8	Trait Anxiety Inventory
9	Social Anxiety Question
10	Negative Life Events in Last Year
11	Antisocial Personality Traits/Conduct Disorder
12	Family Social Support
13	Friend Social Support

In our analysis, we will design cross-cohort and cross-validation schemes to test the logistic regression models at the daily level, and compare them with our proposed algorithms to examine their respective generalizability. We aim to discover new patterns through multivariate data mining techniques, so all available daily and person-level variables were used in our analysis. Day-to-day effect will be examined in our longitudinal analysis rather than average daily effect. Given the very different configurations between our analysis and the previous analysis, findings resulting from these two analyses differ significantly and may not be comparable. For instance, the daily variable "stressfulness today" was reported as a major risk factor in our study but the variable was not used in the prior analysis [Armeli et al. 2010]. Readers can consult Sections 4.1.3 and 4.2.3 for more discussions on findings. We will thus focus on the comparison at the algorithm level between logistic regression and our proposed algorithms.

3. COLLEGE DRINKING DATA

The data used in our analysis was collected from a study completed over a time period of one year at the University of Connecticut Alcohol Research Center. The subject pool consisted of college-aged students enrolled in the Introductory Psychology course at the University of Connecticut who had reported drinking alcohol at least twice in the past month. A survey instrument was designed [Armeli et al. 2010] and was completed by 530 college students in which 52% were female and 86% were Caucasian. Each participant was asked to complete a survey questionnaire approximately one month after starting their school semester. The survey was completed with a distribution of 61% in the fall semester and 39% in the spring semester, forming two cohorts of 323 and 207 participants, respectively.

Participants completed a one-time baseline survey and a 30-day daily diary using a secure Web. Twenty-six students dropped off in the daily reporting phase after completing their baseline survey. The baseline survey questionnaire consisted of over 100 items to measure various person-level measures, including drinking motives, academic performance, personality, recent depression and anxiety symptoms, negative life events, and some demographic items together with drinking behavior of the last month. With the exception of demographic questions and some school status questions, responses to multiple questions in a group were compacted into a risk factor variable with rating scores. Thirteen such composite factor variables were acquired, as listed in Table I. Most of the original questions in the survey took ratings from 1 to 7 scales, with 1 the lowest level and 7 the highest level. For instance, Beck Depression Inventory (brief

Table II. Data Summary

Item	Diary data	Person-level data
Number of records	15120	530
Number of students	504	530
Number of male students	239	254
Number of female students	265	276
Number of covariates	66	22
Number of daytime covariates	43	-
Number of nighttime covariates	13	-
Number of composed covariates	10	-

Table III. Composed Variables

Feature	Definition
depress	daily depressive mood = (sad + dejected) / 2
anxiety	daily anxiety level = (jittery + nervous) / 2
DTSE	(drinking to socialize + drinking to enhance) / 2
DTC × DTSE	drinking to cope × DTSE
depress × DTC	daily depressive mood × DTC
depress × DTSE	daily depressive mood × DTSE
depress × DTC × DTSE	daily depressive mood × DTC × DTSE
anxiety × DTC	daily anxiety level × DTC
anxiety × DTSE	daily anxiety level × DTSE
anxiety × DTC × DTSE	daily anxiety level × DTC * DTSE

Note: DTSE, DTC are two person-level variables and replicated on each day of the 30 days, and all other variables are daily measures, and hence these composed variables are daily-level variables.

version) was calculated by averaging the rating scores for 13 questions asked related to depression, leading to a numerical value ranging from 1 to 7.

Besides the various risk factors, the survey also measured the recent drinking behavior, resulting in four drinking variables. The following three questions were asked in the survey: "the number of occasions in the past 30 days you have consumed an alcoholic drink"; "the number of average drinks when you drink in the past 30 days"; and "the number of occasions you were drunk in the past 30 days". The last drinking variable, Alcohol Dependence Symptoms, is a numerical average of 17 rating questions in the survey including questions like "How often have you experienced blackouts (loss of memory for drinking episodes)?" and "How often have you consumed alcohol instead of eating a meal?" Students who are likely to develop symptoms for alcohol dependence will have a higher value such as 3 (the highest is 5) and students who have little to no symptoms will have a lower value such as 0 or 1.

Then, each day for 30 days, participants accessed a secure website and completed a brief survey between the hours of 2:30 and 7:00 PM. This time window was selected to coincide with most undergraduate students' naturally occurring end of school day, but before they began their activities for that evening. This avoided the potential of reporting under the influence of alcohol. The daily survey questionnaire consisted of 56 items measuring today's daily events, daily stressors, goal progress, current mood states, alcohol-outcome expectancies for tonight, together with the previous night's alcohol use, social interaction, and paid-work/school-work hours. These 56 items corresponded to 56 variables, including 43 daytime and 13 nighttime variables. Table II provides a summary of our data. Particularly, relevant to our analysis, we combined a few daily-level measures and person-level measures according to the suggestions in [Armeli et al. 2010], which formed ten interactive composed measures, as listed in Table III.

Table IV. Summary of Missing Diary Data Entries

Item	Total number	Per student on average
Whole day missed records	4,560	9
Partially missed records	2,878	6
Complete records(male)	3,455	14
Complete records(female)	4,227	16

One of the major issues in human subject studies is the missing data entries. When a student did not report on a specific day, it corresponded to an entire missing record. When a participant answered a question with an invalid response or the data being wrongly input when entering the results into a spreadsheet or database, it corresponded to a record with partially missing entries. The missing data problem was extremely severe for diary data, whereas we only missed 14 entries in the person-level data. As shown in Table IV, for the diary data, records of 9 days were missed entirely on average per student, and records of 6 days had $2 \sim 5$ missing entries. Since little information can be used to impute the entirely missed records, we decided to only impute the missing values for partially incomplete records using Multiple Imputation package [Rubin 1987] provided in SAS [SAS Global Inc. 2010]. After imputation, we had on average 20 days of records for male students and 22 days of records for female students, respectively, to be used in the analysis.

4. OUR APPROACH

A comprehensive description of the proposed approach is given in this section together with comparison results and their interpretation.

4.1. Longitudinal Analysis via Temporally-Correlated SVM

The goal of our longitudinal analysis is to identify risk factors from daily stress, daily negative moods, and drinking expectancies, and to construct a classifier, as a function of the identified daily risk factors, that predicts whether nighttime binge drinking episodes occurr on a specific day. This goal aligns well with previous analysis on the diary data [Armeli et al. 2010, 2000]. However, unlike previous analysis, we do not prespecify the hypothesis which requires manually preselecting only two or three variables for hypothesis test. Variable values will not be averaged over the study period. Instead, we examine within-day effect that relates tonight's drinking to today's observed factors with additional inputs from proximally previous days.

From a machine learning point of view, our task is a variable selection problem. Variable selection methods are often divided along two lines: filter and wrapper methods [Kohavi and John 1997]. The filter approach of selecting variables serves as a preprocessing step to the model construction. The main disadvantage of the filter approach is that it totally ignores the effects of the selected variable subset on the performance of the classification algorithm. The wrapper method searches the optimal variable subsets using the estimated classification accuracy, as the measure of *goodness*, when the subset of variables is used in classification. Thus, the variable selection is being "wrapped around" a particular classification algorithm. Wrapper methods usually outperform filter methods [Guyon and Elisseeff 2003].

Our variable selection method is a wrapper method that is wrapped around a new design of the 1-norm support vector machine (SVM) [Bi et al. 2003; Zuba et al. 2012]. The 1-norm SVM has been known to suppress the number of variables to be used in a predictive model, because the 1-norm regularization enforces vector sparsity. We briefly review the 1-norm SVM in Section 4.1.2. However, the 1-norm SVM does not deal with periodically repeated measurements collected at different time points for related subjects. In fact, most classification approaches assume that sample records

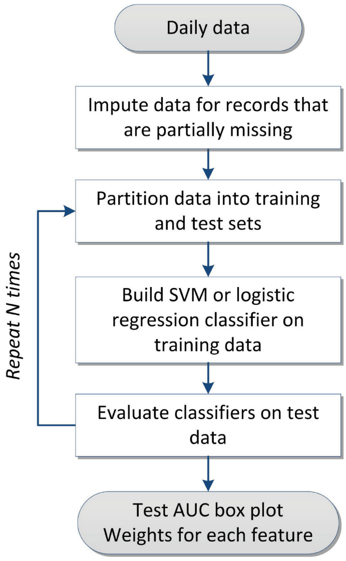
Daily data
Impute data for records that are partially missing
Partition data into training and test sets
Build SVM or logistic regression classifier on training data
Evaluate classifiers on test data
Test AUC box plot
Weights for each feature
Repeat N times

Fig. 2. Flowchart for daily data analysis.

are identically and independently distributed. This assumption is certainly violated in our data. In the diary data, each subject contributed 30 records, and each record corresponded to one of the consecutive 30 days if no data entries were missing. These records were essentially correlated due to the close proximity of the time points at which data was reported by the same person. The most natural correlation for daily longitudinal data is the temporal dependencies. Hence, our variable selection method has to address the selection of risk factors and the correlated records simultaneously. We will devise a so-called temporally correlated SVM in Section 4.1.2 that deals with related records as well as selects variables by imposing the 1-norm regularization.

4.1.1. Flowchart. In our analysis, we first annotated each data record which corresponded to the survey responses reported by a student on a specific day. If a male student had five or more drinks or a female student had four or more drinks at the nighttime of a specific day, this day would be labeled $y = +1$ (a day where binge drinking took place in the evening); otherwise, the day was labeled $y = -1$. We constructed a classifier as a function of as few daily variables as possible to separate days with binge drinking from days without. Figure 2 illustrates the main steps involved in our analysis of diary survey data. After the partially missing records were imputed by the multiple imputation package, the students who missed more than ten entire records were removed from our diary analysis, because day-to-day dynamics could not be sufficiently reflected if a student had very few records. The remaining students were randomly partitioned into training (2/3 of them) and test (1/3 of them) sets. We stratified the partition so that binge drinking days constituted the similar percentage, in both the training and test sets, to that of the full dataset. Classifiers were then built using the diaries in the training set by three methods: the temporally-correlated SVM, standard 1-norm SVM, and logistic regression. The models constructed by these methods were evaluated on the test data. This process was repeated $N = 100$ times by stratified and random split of the full data 100 times, thus forming an effective cross-validation process for validating model generalizability.

Regular logistic regression maximizes likelihood. For a fair comparison, a prior-related term can be added to the logarithm of likelihood with a tuning parameter denoted as c, leading to a maximum-a-posterior version of logistic regression. For all three methods, the tuning parameter c needed to be pre-specified prior to model construction. We used a three-fold cross validation within the training set to adjust the value of this parameter for each partition of the data. As the standard SVM and logistic regression methods were not designed to deal with correlated data records, we added to the data 30 dummy variables to account for the correlation of the 30 days of each student following a similar procedure to the previous analysis [Armeli et al. 2010]. Each of the dummy variables corresponded to one day in the 30-day period and took a value of 1 on that day and 0 on other days to manually create independence between records. Further, it has been found that men and women have different risk factors attributable to their problem drinking behavior. Hence, the process depicted in the flowchart of Figure 2 was repeated separately for male and female students.

4.1.2. Temporally-Correlated Support Vector Machine. SVM is a supervised learning method which has the ability to weigh input factors according to their relevance to the classification target, as determined through the learning process [Bi et al. 2003]. Most SVMs, including the one we implemented in this study, constructs a nonprobabilistic binary linear classifier that predicts whether new records will fall into one category or the other. Our temporally-correlated SVM is based on the sparse formulation of the 1-norm SVM.

The 1-norm SVM constructs a classifier based on a linear function of the form of $\mathbf{w}^T\mathbf{x} + b$, where $\mathbf{w}$ is the weight vector to be determined and $\mathbf{x}$ is the input vector representing one daily record by minimizing the following regularized risk function.

$$\sum_{j=1}^{d} |w_j| + c\sum_{i=1}^{n} \xi_i, \tag{1}$$

where d represents the number of variables/factors in total, n represents the number of records reported by all students in the training set, and $\xi = \max\{0, 1 - y(\mathbf{x}^T\mathbf{w} + b)\}$ denotes the so-called hinge loss [Vapnik 1995], where y represents the class label "binge" versus "non-binge" of an input vector $\mathbf{x}$ which corresponds to a daily record of a student. Since the absolute value of w is used in Eq. (1), it does not correspond to a canonical form of an optimization problem. We hence use the change of variables to convert the problem into the following equivalent optimization problem.

$$\begin{aligned} &\text{Minimize} && \sum_{j=1}^{d} v_j + c\sum_{i=1}^{n} \xi_i \\ &\text{subject to: } y_i\big(\mathbf{w}^T\mathbf{x}_i + b\big) \geq 1 - \xi_i, && \text{where } \xi_i \geq 0,\ i = 1, \ldots, n \\ & && -v_j \leq w_j \leq v_j, \text{ where } v_j \geq 0,\ j = 1, \ldots, d, \end{aligned} \tag{2}$$

where we use v_j to specify the magnitude of w_j, that is, $|w_j| \leq v_j$, which implies $-v_j \leq w_j \leq v_j$. By the change of variables, we replace w by v in the objective function that is to be minimized and require nonnegativity of v associated with each factor. It can be proved that when an optimal solution is found, the optimal value of v is exactly equal to the absolute value of w [Bi et al. 2003]. Minimizing the 1-norm penalty, that is, the first item in the objective function, is known to create a sparse weight vector $\mathbf{w}$. In other words, a large portion of the elements in $\mathbf{w}$ will be driven to 0 at optimality. Hence, only those factors that receive nonzero weights in the linear model $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b$ are selected. Thresholding on the values of $f(\mathbf{x})$ by a cutoff value yields the class labels y for each day of each student. For example, if $f(\mathbf{x}) \geq a$, where a is a discrimination threshold, then on the specific day represented by the vector x, the student had been binge drinking; otherwise, the day belongs to the non-binge class.

Our temporally-correlated SVM is derived based on the 1-norm SVM to deal with data where records are correlated. For the daily data of a specific student, correlation and causal dependence between consecutive days naturally arise. For example, student i reported his daily activities for three consecutive days, corresponding to three records of daily factors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, each of which included the levels of stress, anxiety, or anger mood at the corresponding day, for instance, and three class labels of whether nighttime binge drinking took place y_1, y_2 and y_3. If this student had an increased level of stress on day 1 and its consequences lasted until day 2, the student might have a binge drinking episode on day 2. If the student had a course exam on day 3, it might suppress him from drinking heavily in the evening of day 2. Hence, the outcome of drinking on day 2, y_2, relied not only on $\mathbf{x}_2$, but also on $\mathbf{x}_1$ and $\mathbf{x}_3$. A standard 1-norm SVM or logistic regression constructs a model to predict y_2 based on only $\mathbf{x}_2$ as $y_2 = f(\mathbf{x}_2)$, which does not allow the method to identify the potential effect from the previous day's elevated stress level, for instance. In the temporally-correlated SVM, recorded risk factor values of the proximity days are used to build the predictive model, so y_2 is predicted based on $y_2 = f(\mathbf{x}_2) + \sigma_1 f(\mathbf{x}_1) + \sigma_2 f(\mathbf{x}_3)$, where σ's determine the magnitudes of influence from the proximity days. If the elevated stress value in $\mathbf{x}_1$ causes elevated drinking in y_2, then the cumulative risk from stress can be high over the three days, and the temporally-correlated SVM has the potential of picking up the risk effect.

We associate with the 30 days' records of each student i a parameter matrix $\mathbf{M}$ which is used to specify how each day's drinking outcome is related to the drinking prediction $f(\mathbf{x})$ of the same day and nearby days. We limit our discussion to linear risk functions $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b$. (The discussion here can be extended to nonlinear functions through kernel machines [Vapnik 1998].) Let S_i be the index set for the records of student i, the hinge loss for the standard SVM will be generalized to the set S_i: $Y_i(\mathbf{M}(\mathbf{X}_i\mathbf{w} + b)) \geq 1 - \boldsymbol{\xi}_i$, where $\boldsymbol{\xi}_i \geq 0$, Y_i is a diagonal matrix with diagonal elements equal to y_ℓ, $\ell \in S_i$ $\mathbf{X}_i$ is a matrix with each row equal to $\mathbf{x}_\ell$, $\ell \in S_i$, and $\boldsymbol{\xi}_i$ is a vector containing the hinge loss ξ_ℓ on each day of the student i, that is, $\ell \in S_i$. For each training set in a partition, we solve the following optimization problem for the best within-day effect model $f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b$.

$$\begin{aligned} \text{Minimize} \quad & \sum_{j=1}^{d} v_j + c\sum_{i=1}^{n_s} \mathbf{1}^T\boldsymbol{\xi}_i \\ \text{subject to:} \quad & Y_i(\mathbf{M}(\mathbf{X}_i\mathbf{w} + b)) \geq 1 - \boldsymbol{\xi}_i, \ \text{where } \boldsymbol{\xi}_i \geq 0, \ i = 1, \ldots, n_s, \\ & -v_j \leq w_j \leq v_j, \ \text{where } v_j \geq 0, \ j = 1, \ldots, d, \end{aligned} \tag{3}$$

where n_s is the total number of students in the training set. Notice that the parameter matrix $\mathbf{M}$ is 30 × 30 in size and that its entries need to be estimated from data. However, estimating the full matrix $\mathbf{M}$ will significantly increase the amount of free parameters, in other words, increase model complexity, thus requiring stronger regularization [Vapnik 1995]. One solution we employed in our method was to require $\mathbf{M}$ to be sparse and that each day's drinking outcome relied on the current day t and only three previous days. The magnitude of the influence of these three days on day t's drinking outcome was specified by σ_1, σ_2, and σ_3 for days $t-1$, $t-2$, and $t-3$, respectively. These σ's were estimated from data at the same time as when the model was constructed. We organized the sigma's into a vector $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$. Problem (3) was solved not only for the optimal classifier parameters $(\mathbf{w}, b)$ like in the 1-norm SVM, but also for the best coefficients $\boldsymbol{\sigma}$. Existing solvers for the 1-norm SVM will not be enough to optimize the additional $\boldsymbol{\sigma}$. We derived an efficient alternating optimization algorithm, as depicted in Algorithm 1, to solve Problem (3).

ALGORITHM 1: Solving Temporally-Correlated SVM

```
Input: $\epsilon$ = accuracy tolerance,   $K$ = the maximum number of iterations,
       $\sigma_{init}$ = initial value of $\sigma$,   $v_{max}$ = large machine number
$k = 0$ ;
$o_{old} = v_{max}, o_{new} = v_{max}$;
Initialize $\sigma = \sigma_{init}$;
repeat
    $k = k + 1$;
    $o_{old} = o_{new}$;
    Construct $\mathbf{M}$ with the fixed $\sigma$;
    Solve Problem (3) with $\mathbf{M}$ fixed for the optimal $\mathbf{w}$, $b$;
    Solve Problem (3) with $\mathbf{w}$ and $b$ fixed for the optimal $\sigma$;
    Calculate the objective value of Problem (3) at the above obtained value of $\mathbf{w}$, $b$ and $\sigma$;
    Assign the objective value to $o_{new}$;
until $o_{old} - o_{new} < \epsilon$ or $k \geq K$;
```

The alternating optimization Algorithm 1 starts with an initialization of $\sigma = (\sigma_1,\ \sigma_2,\ \sigma_3)$ to the predefined σ_{init}. Then the algorithm alternates between optimizing the model parameters **w**, b, and the proximal-day influence parameters σ's. It terminates when the objective value of Problem (3) cannot be improved by at least ϵ or the maximum number of iterations is reached.

4.1.3. Results and Discussion. We compared the proposed temporally-correlated SVM, 1-norm SVM, and logistic regression on the analysis of daily dairies of 380 students who reported on 20 or more days during the 30-day study period. In our cross-cohort experiments, the three methods were applied to the cohort recruited during a Fall semester to construct a classifier which was then applied to the cohort recruited during a Spring semester for test. For the temporally-correlated SVM, Algorithm 1 optimized the parameter $\boldsymbol{\sigma} = (0.58,\ 0.68,\ 0.3)$ that measured the magnitude of the influence of the previous three days on the current day. In comparison with the weight of the current day of 1, which was the maximal possible weight σ, the previous two days received rather large weights, 0.58 and 0.68, in the model which reflected their strong influence to the current day's alcohol use behavior. The regularization parameter c was tuned and $c = 0.75$ and $c = 0.63$ were the choice for the temporally-correlated and the standard SVMs, respectively.

Figure 3 shows the *receiver operating characteristic* (ROC) curves of the three classifiers, each trained by one of the methods, when applied to the Spring cohort. An ROC curve is a graphical representation or plot of the true positive rate (sensitivity) versus the false positive rate (1-specificity) for a binary classifier system as its discrimination threshold is varied [Zweig and Campbell 1993]. In a binary classification problem, an example is labeled by class labels of either "positive = binge drinking" or "negative." This results in a total of four possible outcomes. The first case is where the person is predicted to be positive and is in fact positive, known as a true positive. The second case is where the person is predicted to be positive and is in fact negative, known as a false positive. The last two cases are the inverse of the first two; specifically, the person is predicted to be negative and is or is not actually the negative case. These last two cases are known as true negatives or false negatives. In general, a model with good performance should have a ROC curve that is more towards the top left-hand corner,

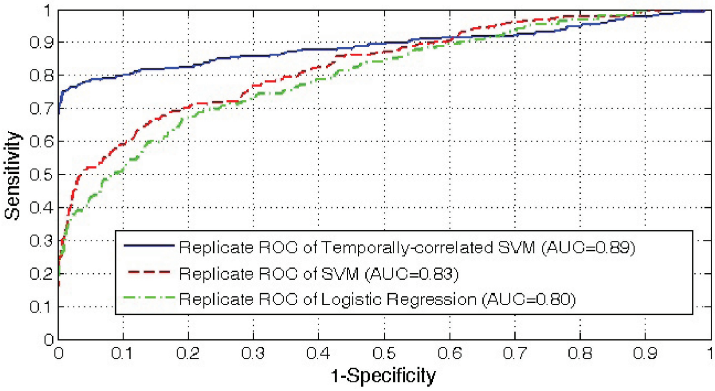

Fig. 3. Test receiver operating characteristic (ROC) curves for binge drinking classifiers. Classifiers were trained on the Fall cohort and tested on the Spring cohort.

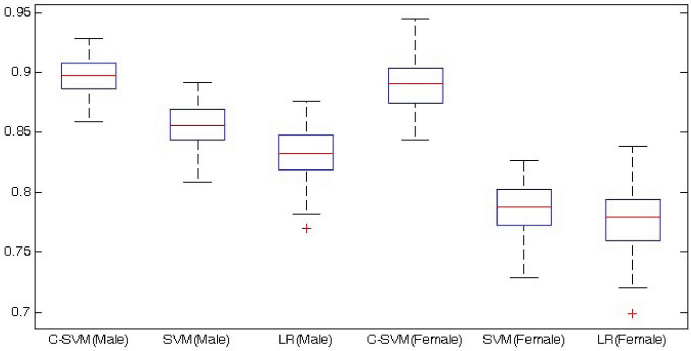

Fig. 4. AUC comparison of temporally-correlated SVM, standard 1-norm SVM, and logistic regression on male and female students.

as this implies higher true positive rates at lower false positive rates. An ROC curve with a straight diagonal line means that the prediction of the classifier is random and therefore not accurate. An ROC curve also comes with a statistic called the *area under the curve* (AUC) which measures the overall performance of a classifier in terms of the probability of the classifier output being correct. From the curves in Figure 3, we can see that the two SVM algorithms outperformed the standard method used in alcohol study literature, logistic regression. The temporally-correlated SVM demonstrated the best performance.

In order to thoroughly compare among the three methods, as discussed in Section 4.1.1, a cross-validation procedure was used to construct classifiers separately for male and female students in 100 trials. During this procedure, Algorithm 1 reported the optimal $\boldsymbol{\sigma} = (0.69,\ 0.72,\ 0.40)$ and $\boldsymbol{\sigma} = (0.63,\ 0.70,\ 0.61)$ for male and female students, respectively, averaged over the 100 trials. Figure 4 presents the AUC bar plots averaged over the 100 models built with, respectively, male and female students. As shown in the figure, SVM models consistently and significantly outperformed logistic regression models, and the temporally-correlated SVM model had the best accuracies. The paired t-test on the 100 pairs of AUC values between the temporally-correlated SVM and the standard SVM achieved p values of 4×10^{-19} and 6×10^{-21} for male and female students, respectively. The paired t-test between the standard SVM and logistic

Table V. Averaged Weight Values of Top 10 Daily Factors in the 100 Temporally-Correlated SVM Models (Male)

Feature	Weight	Percentage(%)	Cumulative percentage(%)
Stressfulness today	2.291	30.0	30.0
DTSE	2.112	27.4	57.4
Negative experiences today-challenging	0.90	11.7	69.1
Daily anxiety level	0.711	9.20	78.3
Hostile	0.251	3.25	81.6
Daytime academic stress	0.133	1.72	83.3
Outcome desirability - feeling careless/irresponsible	0.132	1.71	85.0
Coping - used alcohol to get through it	0.131	1.70	86.7
Sad	0.033	0.43	87.1
Outcome desirability - being clumsy/uncoordinated	0.025	0.32	87.4

Table VI. Averaged Weight Values of Top 10 Daily Factors in the 100 Temporally-Correlated SVM Models (Female)

Feature	Weight	Percentage(%)	Cumulative percentage(%)
Stressfulness today	2.366	41.5	41.5
Coping - used alcohol to get through it	1.162	20.3	61.8
Nighttime used drugs	0.853	14.9	76.7
Daily anxiety level	0.516	9.05	85.75
DTSE	0.324	5.68	91.43
Outcome desirability - being clumsy/uncoordinated	0.222	3.89	95.32
Nighttime overall stressfulness	0.207	3.62	98.94
Excited	0.0201	0.35	99.29
Outcome likely - having good time with friends	0.0174	0.30	99.59
Last night hours of sleep	−0.0164	0.28	99.80

regression achieved p values of 8×10^{-14} and 7×10^{-10} on male and female students, respectively.

The 66 daily risk factors were ranked according to their corresponding weights averaged over the 100 trials in the descending order of the magnitude of the weight. The top ten factors together with the percentage of effects explained by the factors are reported in Table V for male students and Table VI for female students. As designed in the temporarily-correlated SVM, that is, Problem (3), for a student's daily record $\mathbf{x}_t$ on day t, a linear model with parameters $(\mathbf{w}, b)$ is constructed to calculate $\mathbf{x}_t^T\mathbf{w} + b$ as the contribution from day t to y_t, the night time drinking outcome of day t. However, the night time drinking outcome of day t is not only modeled using day t's record, but also the records of the previous three days $t-1$, $t-2$, and $t-3$, by calculating $y_t = (\mathbf{x}_t^T\mathbf{w}+b)+\sigma_1(\mathbf{x}_{t-1}^T\mathbf{w}+b)+\sigma_2(\mathbf{x}_{t-2}^T\mathbf{w}+b)+\sigma_3(\mathbf{x}_{t-3}^T\mathbf{w}+b)$. Notice that the same $\mathbf{w}$ is applied to each of the four daily records. Hence, the weights $\mathbf{w}$ for each daily factor in Tables V and VI were determined under consideration of the proximal influence from three previous days. The model resulted from solving Problem (3) does not identify which specific variables from previous days serve as risk factors for y_t. Instead, it identifies overall important factors when each of the factors has taken a combined value across all the four days in the model.

Based on these two tables, *DTSE* has been consistently chosen as a major stimulator to nighttime binge drinking. This reflects the same observation in the early analysis on this data set [Armeli et al. 2010], but our analysis shows that *DTSE* plays a more significant role in predicting male students' nighttime drinking than female students. Both tables demonstrate that daily *stressfulness* and *anxiety level* are strongly predictive of

Table VII. Drinking Variables

Unique Name	Question/Factor
DrinkFreq	Number of occasions in the past 30 days student has drank
DrinkAmnt	Average number of drinks in the past 30 days per occasion
DrunkFreq	Number of occasions in the past 30 days student was drunk
AlcDep	Alcohol Dependence Symptoms

nighttime binge drinking. We focused on the analysis of daily moderating effect of stress and negative affect. Hence, we included in our models all stress-related and affect-related daily variables in the data, which was distinct from previous work [Armeli et al. 2010], where only averaged daily stress level was used. Besides the *stressfulness* that measured a student's overall stress level, *daytime academic stress* and *nighttime overall stressfulness* were also identified positively associated with nighttime binge drinking for male students and female students, respectively. Additional affect-related variables, *hostile* (explaining 3.25% of the drinking outcome for male), *sad*, and *excited*, were also among the top ten risk factors, but had significantly lower impact than anxiety levels.

A group of eight variables were also acquired to measure the *outcome desirability* and *Outcome likelihood* when students were to drink alcohol tonight. The *outcome desirability - feeling clumsy / uncoordinated* was selected by both the male and female models, but explained less than 5% of the overall drinking effect. The daily survey also asked the approaches students used to deal with their most negative event today, and the *coping method - using alcohol to get through it* was selected as a stimulator to drinking among male students with small effect. Factors explaining over 10% of the drinking outcome included *negative experiences today - challenge* for male students and *nighttime used drugs* for female students. These findings have not been previously reported and require separate replications and further exploration.

4.2. Person-Level Risk Factor Identification via a Composite Learning Scheme

We propose a composite scheme that combines a cluster analysis consecutively with a feature selection to identify person-level risk factors for heavy drinking. The goal here is to recognize different drinking patterns or subtypes based on alcohol use variables and then subsequently identify person-level risk factors that are either associated with a pattern or discriminative between patterns. Section 4.2.1 describes in detail our experimental design.

4.2.1. Flowchart. The 14 missing entries in the person-level data were first imputed using the Multiple Imputaiton Package. Of the total 26 person-level measures, four were variables used to characterize student average drinking behavior, as described in Table VII. These four variables were used to group students into different clusters by the K-Medoids method. We obtained three clusters. According to the values of the four variables averaged within each cluster, we named these clusters 1 - non-risk drinker group; 2 - moderate drinker group; and 3 - heavy drinker group. The individuals in these clusters were then assigned with the labels of 1, 2, 3, respectively.

We built three classifiers, each used to separate one drinker group from the rest. The labeled data was partitioned by gender and then each group was randomly split into a training set containing 2/3 of the subjects in the group and a test set containing the other 1/3 of the subjects. This scheme allows for each classifier to be tested on a set of subjects independent of the set used in classifier training. We stratified the partition so that the ratio of the numbers of students with different labels in the

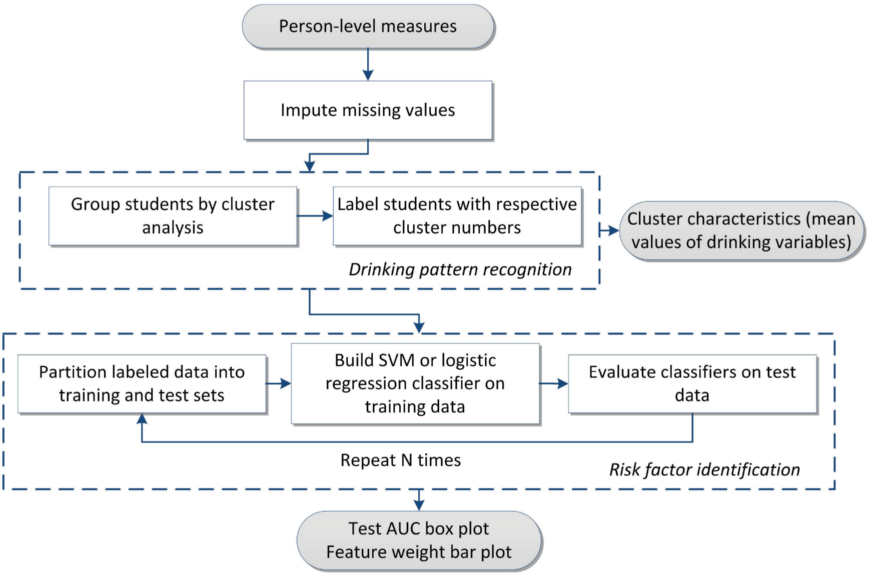
Person-level measures
Impute missing values
Group students by cluster analysis
Label students with respective cluster numbers
Drinking pattern recognition
Cluster characteristics (mean values of drinking variables)
Partition labeled data into training and test sets
Build SVM or logistic regression classifier on training data
Evaluate classifiers on test data
Repeat N times
Risk factor identification
Test AUC box plot
Feature weight bar plot

Fig. 5. Flowchart for person-level data analysis.

partition remained the same as the ratio for the entire data. We repeated the data partition $N = 100$ times. For each partition, classifiers were built using the 1-norm SVM and the logistic regression method based on the training data, and the resultant classifiers were evaluated on the test data. For a fair comparison to SVM, we used the maximum-a-posterior version of logistic regression and required a tuning parameter, similar to c in SVM. Cross validation was used to find a proper c for both methods based on training data only. For each model, we draw the ROC plot with its associated AUC statistic. Therefore, 100 AUC values will be obtained for both SVM and logistic regression models. We average these AUC values into a box plot for the comparison of the two methods (Figure 5).

4.2.2. Cluster Analysis. In the data collected from the baseline survey, each row represented a unique subject or a single completed survey, and each column entry represented a subject's rating value to a risk factor. The data consisted of several attribute types that took categorical, binary, or numerical values. We were able to employ all the variables of different attribute types by applying multiple correspondence analysis (MCA) [Johnson and Wichern 1998]. The MCA technique compacts categorical variables into lower-dimensional space, similar to principle component analysis (PCA) for continuous variables. It operates on an indicator matrix $Z_{n\times k}$, where entries are either 1 or 0, n is the number of samples, and if each variable has k_j categories, k is the sum of k_j over all categorical variables. In general, MCA provides a geometric model of the data and summarizes the relations between the categorized variables [Johnson and Wichern 1998]. This step of preprocessing is more sophisticated and efficient than the typical schemes used in alcohol literature, where a numerical number is simply assigned to each category of a categorical variable.

Additionally, some question items had data that was represented by a 0 or 1, while other questions had data represented as a rank, such as 7 or 5. This could create problems in which some factors would be weighted higher than others because they were represented by a higher numerical value. In order to address this issue, we preformed a

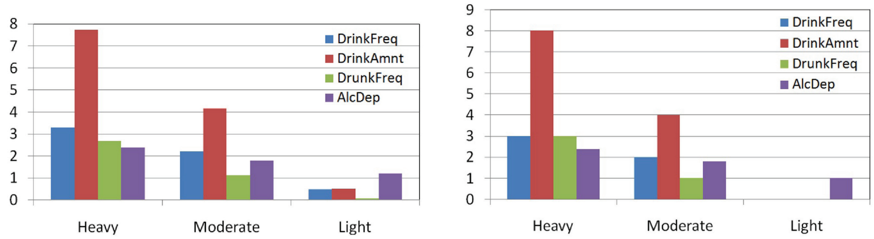

Fig. 6. Drinking behavior clusters: (left) mean values and (right) median values of each drinking variable in Table VII. The "Light" cluster is also named as non-risk cluster. Heavy drinking group (male: 155, female: 62); moderate drinking group (male: 92, female: 151); and non-risk drinking group (male: 41, female: 73).

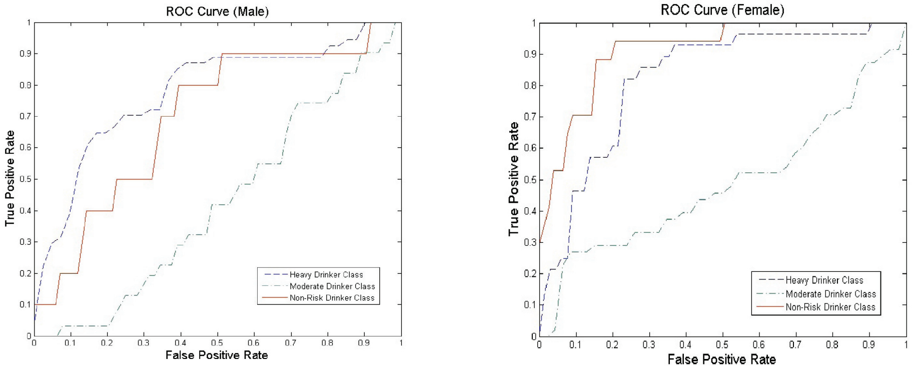

Fig. 7. ROC curves for male drinking classifiers (left) and female drinking classifiers (right). Classifiers were trained on the Fall cohort and tested on the Spring cohort.

standard normalization for each risk factor. Each factor was hence normalized to have a mean value of 0 and a standard deviation of 1.

A K-Medoids clustering algorithm [Reynolds et al. 1992, 2004] was applied to the four variables listed in Table VII with the number of clusters k set to three. Figure 6 shows the mean values and median values of the four variables for each of the three clusters. The median values of each cluster were the cluster medoids (or centers) that the K-Medoids method used to represent individual clusters. Based on the characteristics shown in the two figures, the identified clusters were well separated in terms of the characteristics of drinking behaviors, and we hence named the three clusters heavy drinker, moderate drinker, and non-risk (light) drinker groups.

The K-Medoids algorithm performs cluster analysis that is similar to the commonly used K-Means. K-Medoids is used to divide data into disjoint clusters. This algorithm attempts to minimize a squared error averaged over all clusters. The square error is proportional to the overall distance between the objects in the cluster and an object that is the center or medoid of the cluster [Reynolds et al. 1992]. A medoid is considered to be an object of the cluster whose average dissimilarity to all the objects in the cluster is minimal [Reynolds et al. 1992]. The difference between K-Medoids and K-Means lies in the choice of data objects chosen as centers, and K-Medoids is more robust to noise and outliers.

4.2.3. Results and Discussion. Our initial evaluation of SVM classifiers was done by training a model on the Fall cohort and test the model on the Spring cohort. Figure 7 shows the ROC results of the initial evaluation for males and females classifiers, respectively. We observe that the heavy drinking classifier and non-risk drinking

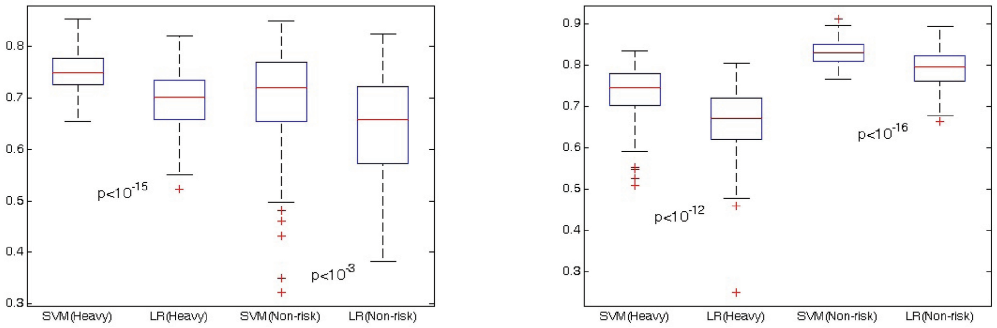

Fig. 8. Average AUC values together with standard deviation bars (box plots) for male drinking classifiers (left) and female drinking classifiers (right) obtained by SVM and logistic regression (LR).

classifier for female performed the best among all models. The heavy drinking and non-risk drinking male classifiers were moderately accurate. The moderate drinker classifiers for both male and female performed poorly and were closer to a random-guessing classifier. However, this is expected, as there is a lack of distinguishing features that could appropriately differentiate between the middle class–moderate drinkers and the extreme cases at the two opposite ends: heavy drinkers and light drinkers. The same was observed for the logistic regression models (figures not shown) as well. Hence, we excluded the moderate drinker classifier from further analysis.

To compare between the 1-norm SVM and the logistic regression that were both applied to the same data partitions, we plot the averaged AUC values in Figure 8. The box plots of AUC were obtained by calculating the average and standard deviation of the AUC evaluation of the 100 data partitions, which resulted in 100 SVM models and logistic regression models built on the male population and female population, respectively. Two sets of the models were built: one for distinguishing heavy drinkers from the others and the other for discriminating non-risk drinkers from the others. We can see from Figure 8 that the accuracies of SVM models were significantly better than those of logistic regression models according to the paired t-test on the 100 pairs of AUC values. The performance differed significantly between SVM and logistic regression with p values of 1.51×10^{-16} (male students, heavy drinker classifier), 2.23×10^{-4} (male students, non-risk drinker classifier), 1.29×10^{-13} (female students, heavy drinker classifier), and 3.58×10^{-17} (female students, non-risk drinker classifier), respectively.

The SVM training process yields a linear classifier with weights associated with each of the risk factors. These weights specify whether a specific factor is a motivator for drinking heavily (positive weight), whether a factor has no effect (a weight value of 0), or whether the factor is a deterrent for preventing from drinking too much (negative weight). Figure 9 shows the weight values for various factors in the different classifiers averaged across the 100 trials.

As shown in Figure 9 (the two plots on the left), *drinking to enhance* was an important motivator to heavy drinking behavior for both male (ranked as the first motivator) and female students (ranked as the third motivator), whereas *drinking to conform* was their common deterrent. Other than *drinking to enhance*, *antisocial personality* also tended to motivate male heavy drinkers to drink heavily. For female heavy drinkers, the top two motivators for their heavy drinking behavior were *drinking to socialize* and *sensation seeking*. The result also shows that older female heavy drinkers tended to drink less.

Based on the non-risk drinker classifiers for both male and female students, *drinking to enhance* also motivated the non-risk drinkers to drink more, as shown in Figure 9

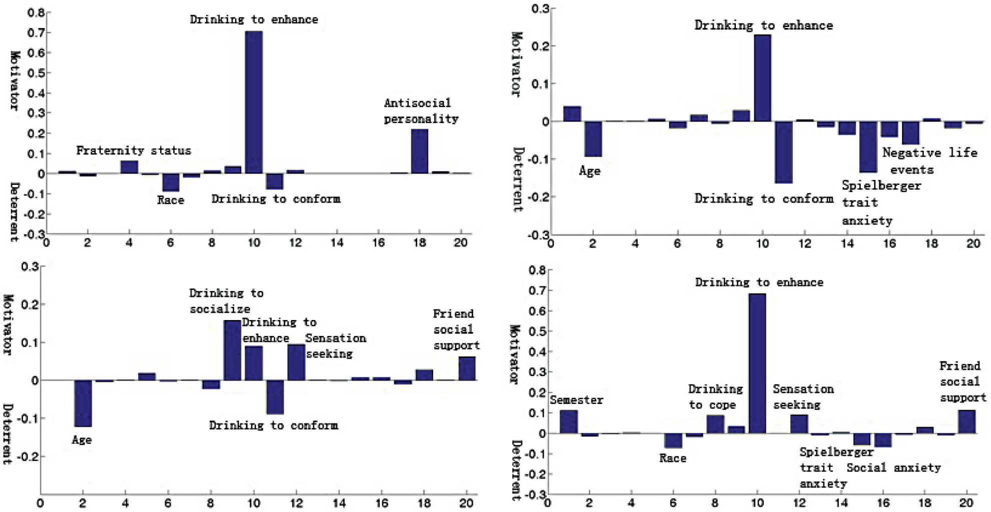

Fig. 9. Averaged weight values for various factors in the classifiers that were built for (top-left) male heavy drinkers, (top-right) male non-risk drinkers, (bottom-left) female heavy drinkers, and (bottom-right) female non-risk drinkers.

(the two plots on the right). It is interesting to observe that male students who were low-risk or non-alcohol users tended to regulate their drinking at elevated negative life events if they also had relatively high *drinking to conform* motive and *trait anxiety index* value. Older age was another important regulator for male non-risk alcohol users to drink less. For female students, *Spielberger trait anxiety* and *social anxiety* were the top factors to protect non-risk drinkers. Besides *drinking to enhance*, *drinking to cope*, *sensation seeking*, and *friend social activity* tended to stimulate female students who were low alcohol users to move towards heavier users. Our female non-risk alcohol user group resulting from the data-driven cluster analysis were not balanced between the two semesters with more students recruited in the Spring semester, which is shown in Figure 9 (bottom-right).

Early work [Armeli et al. 2010] used only DrinkFreq and DrinkAmnt in Table VII separately as the dependent variable, and performed two analyses: logistic regresion for DrinkFreq and linear regression for DrinkAmnt, respectively. Thus, each analysis examined only one drinking-related measure and did not capture how heavily an individual drank in comprehensive characteristics. Instead, our method constituted a cluster analysis which was applied to all four drinking-related variables in the data and created clusters that distinguished on all these measures. Some of our results are consistent to the early analysis, but many of our findings have not been reported before and may bring insights that benefit from further and thorough investigations.

5. CONCLUSION AND DISCUSSION

In this article, we have presented two machine learning methods for a secondary analysis of the data collected in a college drinking study. The proposed longitudinal analysis method can effectively handle day-to-day dynamics and identify daily risk factors for binge drinking under consideration of the proximal influence from nearby days, which advances the state of the art. The main daily factors associated with binge drinking that were identified by our approach enclosed daily stress and anxiety levels, and thus may help to advance the understanding of stress- and negative affect-related drinking (SNAD) behavior [Simons et al. 2005].

The proposed composite machine learning method that associates risk factors with heavy drinking behavior first empirically defines subtypes of drinking behavior by cluster analysis and then detects subtype-related risk factors. This method was applied to the analysis of person-level measures of students (the initial survey) in the present study. However, the method is generally feasible for use in analysis of daily diaries. By choosing sparse modeling techniques, such as the 1-norm support vector machine, we form models that select the most important risk factors and quantify the interplay of these factors. The discovered person-level risk factors and the interaction of these factors may offer insights to new designs of college alcohol interventions.

By comparing to traditional analytics, we have demonstrated the effectiveness of the proposed methods and the differences in the kind of findings they can find. Enclosing separate sections of results and discussion for each of the two proposed methods allows us to discuss more thoroughly the strength of the algorithms, results, and comparisons with respective algorithms. Our approach constructs classifiers that achieve better generalization performance, in other words, better prediction accuracy, than logistic regression, as shown by the included ROC plots and AUC box plots. Better model generalizability predicts that the model can perform better on future data sets. Given the more accurate prediction of our models on new cases that are not used in the construction of the model, the risk factors identified by our models are more likely to be valid. We expect that our methodology can help to analyze other dairy data.

Our work has a number of limitations. Although our method, as a data-driven approach, has potential to discover new knowledge from large quantities of data, without any guidance from psychological and medical theories, it may find patterns biased to the empirical sample. An ideal scenario is to combine both the carefully-collected data and the cumulated human knowledge for discovery of new knowledge. Development of machine learning algorithms that can model data with guidance from domain knowledge might further enhance the knowledge discovery process.

Much of our analysis was based on support vector machines. Although support vector machines have a variety of advantages [Vapnik 1998] and the 1-norm support vector machine can automatically select the most relevant variables, other classification methods, such as Bayesian nets or graphical models, may be effective and valuable alternatives for improving generalizability. The proposed composite machine learning scheme is suitable for use in conjunction with these alternatives as well.

The temporally-correlated support vector machine uses a sparse matrix $\mathbf{M}$ to link the influence of different nearby days when predicting today's outcome. Our current implementation of $\mathbf{M}$ based on only previous three days may limit us from identifying long-term cumulative risk. There may exist other designs that do not rely on $\mathbf{M}$. For example, building a classifier that predicts today's drinking based on all the factors observed on related days may help to pinpoint which factors from which nearby day influence today's drinking.

ACKNOWLEDGMENTS

The authors would like to thank Dr. V. Hesselbrock at UConn Alcohol Research Center for insightful discussions and specifications about the application scenario of the proposed techniques. Mr. M. Zuba and J. Gilbert, as students in Dr. J. Bi's class, were involved in the early analysis of a small portion of the current data and provided some useful initial feedback to motivate the current study.

REFERENCES

Abar, C. C. and Maggs, J. L. 2010. Social influence and selection processes as predictors of normative perceptions and alcohol use across the transition to college. *J. College Student Develop. 51*, 5, 496–508.

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders : DSM-IV*. American Psychiatric Association, Washington D.C.

Armeli, S., Carney, M. A., Tennen, H., Affleck, G., and O'Neil, T. 2000. Stress and alcohol use: A daily process examination of the stressor-vulnerability model. *J. Personality Social Psychol. 78*, 5, 979–994.

Armeli, S., Conner, T. S., Cullum, J., and Tennen, H. 2010. A longitudinal analysis of drinking motives moderating the negative affect-drinking association among college students. *Psychol. Addictive Behav. 24*, 1, 38–47.

Armeli, S., Tennen, H., Todd, M., Carney, M. A., Mohr, C., Affleck, G., and Hromi, A. 2003. A daily process examination of the stress-response dampening effects of alcohol consumption. *Psychol. Addictive Behav. 17*, 4, 266–276. ID: 4642330957.

Auerbach, K. J. and Collins, L. M. 2006. A multidimensional developmental model of alcohol use during emerging adulthood. *J. Studies Alcohol 67*, 6, 917–925.

Beck, K. H., Arria, A. M., Caldeira, K. M., Vincent, K. B., O'Grady, K. E., and Wish, E. D. 2008. Social context of drinking and alcohol problems among college students. *Am. J. Health Behav. 32*, 4, 420–430.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. 2003. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res. 3*, 1229–1243.

Carter, A. C., Brandon, K. O., and Goldman, M. S. 2010. The college and non-college experience: A review of the factors that influence drinking behavior in young adulthood. *J. Studies Alcohol Drugs 71*, 5, 742–750.

Chung, T., Martin, C. S., and Winters, K. C. 2005. Diagnosis, course, and assessment of alcohol abuse and dependence in adolescents. In *Recent Developments in Alcoholism*, vol. 17, Springer, New York, 5–27. ID: 111389880.

DeHart, T., Tennen, H., Armeli, S., Todd, M., and Mohr, C. 2009. A diary study of implicit self-esteem, interpersonal interactions and alcohol consumption in college students. *J. Exp. Social Psychol. 45*, 4, 720–730.

DeJong, W., Schneider, S. K., Towvim, L. G., Murphy, M. J., Doerr, E. E., and Simonsen, N. R. 2009. A multisite randomized trial of social norms marketing campaigns to reduce college student drinking: A replication failure. *Subst. Abuse 30*, 2, 127–140.

Donovan, J. E., Jessor, R., and Jessor, L. 1983. Problem drinking in adolescence and young adulthood. A follow-up study. *J. Stud. Alcohol 44*, 1, 109–37.

Guyon, I. and Elisseeff, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, NY.

Hingson, R. W., Heeren, T., Zakocs, R. C., Kopstein, A., and Wechsler, H. 2002. Magnitude of alcohol-related mortality and morbidity among U.S. college students ages 18–24. *J. Stud. Alcohol 63*, 136–144.

Hingson, R. W., Wenxing, Z., and Weitzman, E. R. 2009. Magnitude of and trends in alcohol-related mortality and morbidity among U.S. college students ages 18–24: Changes from 1998 to 2005. *J. Stud. Alcohol Drugs Suppl. 16*, 12–20.

International Center for Alcohol Policies. 2003. Policy Issues. http://www.icap.org/PolicyTools /ICAPBlueBook/BlueBookModules/6BingeDrinking/tabid/167/Default.aspx#1.

Jackson, K. M. and Sher, K. J. 2005. Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychol. Addict. Behav. 19*, 4, 339–51.

Jellinek, E. M. 1960. *The Disease Concept of Alcoholism*. Hilhouse, New Brunswick, NJ.

Johnson, R. A. and Wichern, D. W. 1998. *Applied Multivariate Statistical Analysis* 4th Ed. Upper Saddle River, NJ.

Kohavi, R. and John, G. H. 1997. Wrappers for feature subset selection. *Artif. Intell. 97*, 1–2, 273.

LaBrie, J. W., Migliuri, S., Kenney, S. R., and Lac, A. 2010. Family history of alcohol abuse associated with problematic drinking among college students. *Addict. Behav. 35*, 7, 726–729.

Lewis, K. P., Lindgren, M. A., Fossos, N., Neighbors, L., and Oster-Aaland, C. 2009. Examining the relationship between typical drinking behavior and 21st birthday drinking behavior among college students: Implications for event-specific prevention. *Addiction 104*, 5, 760–767.

Maggs, J. L. and Schulenberg, J. E. 2005. Trajectories of alcohol use during the transition to adulthood. *Alcohol Res. Health 28*, 4, 195–201.

Moulton, M., Moulton, P., Whittington, A. N., and Cosio, D. 2000. The relationship between negative consequence drinking, gender, athletic participation, and social expectancies among adolescents. *J. Alcohol Drug Educa. 45*, 2, 12–22.

National Institute of Alcohol Abuse and Alcoholism. 2002. A call to action: Changing the culture of drinking at U.S. colleges. In *Task Force of the National Advisory Council on Alcohol Abuse and Alcoholism*. U.S. Department of Health and Human Services. http://www.collegedrinkingprevention.gov/.

NATIONAL INSTITUTE ON DRUG ABUSE. 1999. Monitoring the future, national survey results on drug use.

NATIONAL INSTITUTES OF HEALTH. 2008. Secondary analysis of existing alcohol epidemiology data. http://grants.nih.gov/grants/guide/pa-files/PA-08-167.html.

OFFICE OF JUVENILE JUSTICE AND DELINQUENCY PREVENTION. 2005. Drinking in America: Myths, Realities, and Prevention Policy. Tech. rep. U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention, Washington, D.C.

O'MALLEY, P. M. AND JOHNSTON, L. D. 2002. Epidemiology of alcohol and other drug use among american college students. *J. Stud. Alcohol Suppl. 14*, 23–39.

O'ONNOR, R. M. AND COLDER, C. R. 2005. Predicting alcohol patterns in first year college students through motivational systems and reasons for drinking. *Psychol. Addict. Behav. 19*, 1, 10–20.

PATRICK, M. E. AND LEE, C. M. 2010. Comparing numbers of drinks: College students' reports from retrospective summary, followback, and prospective diary measures. *J. Stud. Alcohol Drugs 71*, 4, 554–561.

PERKINS, H. W. 1999. Stress-motivated drinking in collegiate and postcollegiate young adulthood: Life course and gender patterns. *J. Stud. Alcohol 60*, 2, 219–27.

REED, M. B., WANG, R., SHILLINGTON, J. D., CLAPP, A. M., AND LANGE, J. E. 2007. The relationship between alcohol use and cigarette smoking in a sample of undergraduate college students. *Addict. Behav. 32*, 3, 449–464.

REYNOLDS, A. P., RICHARDS, G., DE LA IGLESIA, B., AND RAYWARD-SMITH, V. J. 1992. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *J. Math. Mode. Algor. 5*, 4, 475–504.

REYNOLDS, A. P., RICHARDS, G., AND RAYWARD-SMITH, V. J. 2004. The application of k-medoids and pam to the clustering of rules. In *Intelligent Data Engineering and Automated Learning,* Lecture Notes in Computer Science, vol. 3177, Springer Verlag, Berlin Heidelberg, 173–178.

ROOM, R. 1996. Patterns of family responses to alcohol and tobacco problems. *Drug Alcohol Rev. 15*, 2, 171–181.

RUBIN, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, NY.

SAS GLOBAL INC. 2010. *SAS Manual, version 9.2*.

SHER, K. J., JACKSON, K. M., AND STEINLEY, D. 2011. Alcohol use trajectories and the ubiquitous cat's cradle: Cause for concern? *J. Abnormal Psychol. 120*, 2, 322–335.

SIMONS, J. S., GAHER, R. M., OLIVER, M. N., BUSH, J. A., AND PALMER, M. A. 2005. An experience sampling study of associations between affect and alcohol use and problems among college students. *J. Stud. Alcohol 66*, 4, 459–69. ID: 106180217.

SINGER, J. D. AND WILLETT, J. B. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, Oxford, U.K.

SINGLETON, R. A. AND WOLFSON, A. R. 2009. Alcohol consumption, sleep, and academic performance among college students. *J. Stud. Alcohol Drugs 70*, 3, 355–363.

STAPPENBECK, C. AND FROMME, K. 2010. A longitudinal investigation of heavy drinking and physical dating violence in men and women. *Addict. Behav. 35*, 5, 479–485.

STEWART, S. H., MORRIS, E., MELLINGS, T., AND KOMAR, J. 2006. Relations of social anxiety variables to drinking motives, drinking quantity and frequency, and alcohol-related problems in undergraduates. *J. Mental Health 15*, 6, 671–682.

STOLLE, M., SACK, P. M., AND THOMASIUS, R. 2009. Binge drinking in childhood and adolescence: Epidemiology, consequences, and interventions. *Deutsches Rzteblatt Int. 106*, 19, 323–8.

TALBOTT, L. L., UMSTATTD, M. R., USDAN, S. L., MARTIN, R. J., AND GEIGER, B. F. 2010. Validation of the drinking context scale (DCS-9) for use with non-adjudicated first year college students. *Addict. Behav. 35*, 5, 510–512.

TENNEN, H., AFFLECK, G., ARMELI, S., AND CARNEY, M. A. 2000. A daily process approach to coping: Linking theory, research, and practice. *Am. Psychol. 55*, 6, 626–36. ID: 119360046.

TIMBERLAKE, D. S., HOPFER, C. J., RHEE, S. H., FRIEDMAN, N. P., HABERSTICK, B. C., LESSEM, J. M., AND HEWITT, J. K. 2007. College attendance and its effect on drinking behaviors in a longitudinal study of adolescents. *Alcohol. Clinic. Exp. Res. 31*, 6, 1020–1030.

TOUMBOUROU, J. W., HEMPHILL, S. A., MCMORRIS, B. J., CATALANO, R. F., AND PATTON, G. C. 2009. Alcohol use and related harms in school students in the USA and Australia. *Health Promo. Int. 24*, 4, 373–382.

TURNER, J., KELLER, A., AND BAUERLE, J. 2010. The longitudinal pattern of alcohol-related injury in a college population: Emergency department data compared to self-reported. *Am. J. Drug Alcohol Abuse 36*, 4, 194–198.

VAPNIK, V. 1995. *The Nature of Statistical Learning Theory*. Springer, New York, NY.

VAPNIK, V. 1998. *Statistical Learning Theory*. John Willey & Sons, Inc, New York, NY.

Walls, T. A., Fairlie, A. M., and Wood, M. D. 2009. Parents do matter: A longitudinal two-part mixed model of early college alcohol participation and intensity. *Journal of Studies on Alcohol and Drugs 70*, 6, 908–918.

Wechsler, H. and Austin, S. B. 1998. Binge drinking: The five/four measure. *J. Stud. Alcohol 59*, 1, 122–4.

Wechsler, H., Dowdall, G. W., Davenport, A., and Castillo, S. 1995a. Correlates of college student binge drinking. *Am. J. Public Health 85*, 7, 921–6.

Wechsler, H., Moeykens, B., Davenport, A., Castillo, S., and Hansen, J. 1995b. The adverse impact of heavy episodic drinkers on other college students. *J. Stud. Alcohol 56*, 6, 628–634. ID: 121443215.

Wechsler, H. and Nelson, T. F. 2001. Binge drinking and the American college student: What's five drinks? *Psychol. Addict. Behav. : J. Soc. Psychol. Addict. Behav. 15*, 4, 287–91.

Wood, M. D., Fairlie, A. M., Fernandez, A. C., Borsari, B., Capone, C., Laforge, R., and Carmona-Barros, R. 2010. Brief motivational and parent interventions for college students: A randomized factorial study. *J. Consult. Clinic. Psychol. 78*, 3, 349–361.

Zuba, M., Gilbert, J., Wu, Y., Bi, J., Tennen, H., and Armeli, S. 2012. 1-norm support vector machine for college drinking risk factor identification. In *Proceedings of the ACM SIGHIT International Health Informatics Symposium*. 112–121.

Zucker, R. A., Fitzgerald, H. E., and Moses, H. D. 1995. Emergence of alcohol problems and the several alcoholisms: A developmental perspective on etilogic theory and life course trajectory. In *Developmental Psychopathology*, *Vol. 2, Risk, Disorder, and Adaptation*, D. Cicchetti and D. J. Cohen, Eds., 677–711.

Zweig, M. H. and Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem. 39*, 4, 561–577.

Received November 2012; revised March 2013; accepted April 2013