ORIGINAL RESEARCH

A Machine Learning Approach to Identification of Unhealthy Drinking

Levi N. Bonnell, MPH, Benjamin Littenberg, MD, Safwan R. Wshah, PhD, and Gail L. Rose, PhD

Introduction: **Unhealthy drinking is prevalent in the United States, and yet it is underidentified and undertreated. Identifying unhealthy drinkers can be time-consuming and uncomfortable for primary care providers. An automated rule for identification would focus attention on patients most likely to need care and, therefore, increase efficiency and effectiveness. The objective of this study was to build a clinical prediction tool for unhealthy drinking based on routinely available demographic and laboratory data.**

Methods: **We obtained 38 demographic and laboratory variables from the National Health and Nutrition Examination Survey (1999 to 2016) on 43,545 nationally representative adults who had information on alcohol use available as a reference standard. Logistic regression, support vector machines, k-nearest neighbor, neural networks, decision trees, and random forests were used to build clinical prediction models. The model with the largest area under the receiver operator curve was selected to build the prediction tool.**

Results: **A random forest model with 15 variables produced the largest area under the receiver operator curve (0.78) in the test set. The most influential predictors were age, current smoker, hemoglobin, sex, and high-density lipoprotein. The optimum operating point had a sensitivity of 0.50, specificity of 0.86, positive predictive value of 0.55, and negative predictive value of 0.83. Application of the tool resulted in a much smaller target sample (75% reduced).**

Conclusion: **Using commonly available data, a decision tool can identify a subset of patients who seem to warrant clinical attention for unhealthy drinking, potentially increasing the efficiency and reach of screening. (J Am Board Fam Med 2020;33:397–406.)**

Keywords: **Alcohol Drinking, Alcoholism, Area Under Curve, Clinical Decision Rules, Decision Trees, Logistic Models, Machine Learning, Neural Networks (Computer), Nutrition Surveys, Support Vector Machine**

Introduction

An estimated 27% of adults in the United States drink alcohol at a level considered unhealthy,[1] which is defined as consuming ≥1 drink per day for women or ≥2 for men or binge drinking (consuming ≥4 drinks on the same occasion for women or ≥5 for men) at least once in the past year.[2] Consuming more than the recommended amount of alcohol is a major risk factor for health and social issues, injuries, accidents, and early death.[3–5] Unhealthy drinking has been associated with cancer, pancreatitis, liver disease, psychopathology, sleep problems, hypertension, and other serious diseases,[6–10] costing the United States $249 billion in 2010.[11] Moreover, 88,000 deaths are attributable to consuming unhealthy levels of alcohol each year,[12] making it the third leading preventable cause of death in the United States behind tobacco use and poor diet/lack of exercise.

The United States Preventive Services Task Force recommends screening for unhealthy drinking among adults ages 18 and older,[13] and valid screening tools such as the Alcohol Use Disorders Identification Test (AUDIT),[14] AUDIT-

This article was externally peer reviewed.

Submitted 15 November 2019; revised 31 January 2020; accepted 5 February 2020.

From University of Vermont College of Medicine, Burlington (LNB, BL, GLR); University of Vermont, College of Engineering and Mathematical Sciences, Burlington (SRW).

Funding: This work was supported by the National Institute of Alcohol Abuse and Alcoholism award number 1R41AA025297 to Gail L. Rose (PI).

Conflicts of Interest: none.

Corresponding author: Levi Bonnell, MPH, 89 Beaumont Ave S459, Burlington, VT, 04505 (E-mail: levi.bonnell@med.uvm.edu)

Consumption,[15] and the Single Alcohol Screening Question[16] exist for this purpose.

Primary Care Providers (PCPs) have an important role in identifying people with unhealthy drinking; yet, screening rates in primary care are low. In a representative survey of the US population, only 25% reported having been screened for alcohol use in the last year.[17] Barriers to screening include lack of time and administrative support, need for modifications to office workflow, lack of training for PCPs, the stigma associated with alcohol misuse, and the fact that universal screening will not be applicable to the majority of patients.[18–20] Efforts to impose universal screening through the use of electronic clinical reminders and/or performance measures have improved screening rates in some health care systems but are inconsistently used and can be hampered by low clinical staff buy-in.[21,22]

An alternative approach is a clinical prediction rule, which can automatically identify patients most likely to have unhealthy drinking, thereby reducing the burden on PCPs and staff. Previous research has shown that clinical prediction rules using prospectively collected data can successfully identify unhealthy drinking. Hartzet al[23] used logistic regression and 40 laboratory values to distinguish 426 heavy drinkers from 188 light drinkers. Lichtensteinet al[24] used linear regression plus clinical and laboratory values to predict heavy drinking. Harasymiwet al[25,26] used discriminant function analysis to predict patient-reported alcohol use from a set of blood chemistry profiles. Korzec and colleagues[27] built a predictive test for unhealthy drinking based on laboratory values and a clinical questionnaire using Bayesian networks. However, the generalizability of these studies is limited by small sample sizes and highly selected populations. Furthermore, questionnaires or prospective data collection offer little advantage over universal screening. Finally, neither logistic regression nor discriminant function analysis accommodate missing values, which are common in clinical data.

Clinical prediction rules using large, existing datasets and machine learning methods are gaining momentum in the medical literature and have been used to predict poststroke mortality,[28] in-hospital mortality,[29] peripheral artery disease and future mortality risk,[30] infection in the emergency department,[31] and mortality among colon cancer patients,[32] to mention a few.

The purpose of this study was to build a clinical prediction rule for unhealthy drinking based on routinely collected demographic, clinical, and laboratory data and to compare its performance to a universal screening strategy. We hypothesized that a clinical prediction rule could discriminate patients with greater likelihood of unhealthy drinking from those with a low probability of unhealthy drinking who would not require further evaluation. The population of patients needing further evaluation would, therefore, be smaller and have a higher prevalence of unhealthy drinking and have a greater yield from additional evaluations. In this way, a prediction rule could save time and clinical resources, relieving providers from a function that is challenging to implement reliably.[16,18,19,33]

Materials and Methods

Data Source

Ideally, a clinical prediction model should be developed in the context in which it is intended to be used, based on data available in that context. However, drinking data are inconsistently recorded in electronic health records (EHRs). Therefore, to test our hypothesis that a machine learning approach could be used to build a model for identifying unhealthy drinking, we used a dataset that reliably collected drinking data from each patient.

We obtained deidentified demographic, clinical, and laboratory information on 43,545 nationally representative adults from the National Health and Nutrition Examination Survey (NHANES) from 1999 to 2016. To be included, the records needed responses to the alcohol questions to be used as a reference standard. Individuals younger than 18 years did not receive these questions. Demographic and clinical variables included age, sex, smoking status, height, weight, systolic and diastolic blood pressure, and resting heart rate. Laboratory data included 30 variables from routine clinical chemistries and hemograms (see Table 1). These variables were selected based on prior literature, clinical judgment, and the likelihood that the candidate predictor would be available in routine medical records.[23,34,35] Drinking data were used to classify patients as having either unhealthy drinking or low-risk drinking. Unhealthy drinking was defined by ≥1 drink per day for women or ≥2 for men or binge drinking ≥1 per month in the past 12 months (≥4 drinks on the same occasion for women or ≥5 for

Table 1. Characteristics of the Cohort Stratified by Unhealthy Drinking Status

Demographic Information	Reference Group: Unhealthy Drinkers (n = 11,464), % or Median	Reference Group: Low-Risk Drinkers (n = 32,081), % or Median	*P* Value*
Sex, male†	67%	42%	<0.001
Smoking, current†	36%	15%	<0.001
Age, years†	38	53	<0.001
Clinical Information			
Height, cm	171.2	165.4	<0.001
Weight, kg	80	77.3	<0.001
Systolic blood pressure, mm Hg†	120	122	<0.001
Diastolic blood pressure, mm Hg	72	70	<0.001
Resting pulse rate, 60-second count	72	72	0.13
Chemistry			
Calcium, mg/dL	9.4	9.4	<0.001
Chloride, mmol/L	104	104	<0.001
Phosphorus, mg/dL	3.7	3.7	<0.001
Potassium, mmol/L	4	4	0.006
Sodium, mmol/L	139	139	<0.001
Blood urea nitrogen, mmol/L†	4.3	4.6	<0.001
Creatinine, mg/dL†	0.86	0.82	<0.001
Bicarbonate, mmol/L	25	25	<0.001
Glucose, mg/dL	90	93	<0.001
Uric acid, mg/dL†	5.6	5.2	<0.001
Serum osmolality, mOsm/kg	277	278	<0.001
Liver function			
Bilirubin, mg/dL	0.7	0.6	<0.001
Alanine aminotransferase, U/L	23	20	<0.001
Aspartate aminotransferase, U/L	24	23	<0.001
Alkaline phosphatase, U/L	65	68	<0.001
Gamma-glutamyl transpeptidase, U/L†	23	19	<0.001
Lactate dehydrogenase, U/L†	124	130	<0.001
Protein, g/dL	7.2	7.2	<0.001
Albumin, g/L†	44	42	<0.001
Hematology			
Hemoglobin, g/dL†	14.8	13.9	<0.001
Hematocrit, %†	43.4	41.2	<0.001
Mean corpuscular volume, fL†	90.5	89.8	<0.001
Mean cellular hemoglobin, pg†	30.9	30.5	<0.001
Red blood cell distribution width, %	12.6	12.9	<0.001
White blood cell count, 1000/μL	7.1	6.9	<0.001
Platelet count, 1000/μL	8.1	8.1	<0.001
Lipids			
Total cholesterol, mg/dL	194	193	0.25
High density lipoprotein, mg/dL†	51	50	<0.001
Calculated low density lipoprotein, mg/dL	110	111.4	0.002
Triglyceride, mg/dL	118	121	0.18

*P value determined by χ^2 or Wilcoxon rank-sum test. Because the rank-sum tests considers the entire distribution of each group, it can detect statistically significant differences even when the medians are identical.
†Variables included in final prediction model.

men). Individuals not meeting criteria for unhealthy drinking were classified as low risk. This category includes nondrinkers.

The data were randomly split into 3 independent sets: a training set (65%) for initial development of the model, a validation set (15%) to evaluate the initial model, and a test set (20%) to determine the final fit of the model to the data. The test set was stored separately until a final prediction algorithm was created and ready to use. Univariate analyses were performed to ensure the 3 random subsets were similar.

Model Development and Selection

Six candidate machine learning methods were evaluated to determine the most appropriate approach to use for building a clinical prediction rule with this dataset. Logistic regression,[36] support vector machines,[37] neural networks,[38] k-nearest neighbors,[39] decision trees,[40] and random forests[41] were used individually to create clinical prediction rules for unhealthy drinking using the training set. These methods were chosen based on prior literature[42,43] and because they each have unique advantages and disadvantages for classification (Appendix). Each method was tuned to maximize prediction in the training data using all 38 variables. The decision tree and random forest methods used techniques to extract information from missing values. Essentially, missing data were counted as another level or value of the variable. All resulting clinical prediction rules were run against the validation dataset, and the 1 with the largest area under the receiver operating characteristic curve (AUC)[44] (the random forest) was selected as the target for further evaluation. Variables with an information gain of less than 2% (a measure of importance of each variable in predicting unhealthy drinking) were removed to create a more parsimonious and reproducible clinical prediction rule.[45]

Model Performance

We calculated the performance of the clinical prediction rule in the test set at various thresholds (estimated probabilities of unhealthy drinking), forming a receiver operating characteristic curve. Performance parameters included accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and workload improvement ("savings"). An operating threshold was chosen to optimize these values, with priority given to specificity over sensitivity. Accuracy was calculated as the number of correctly classified patients (true positives + true negatives) divided by the total population. The improvement in screening workload attributable to the clinical prediction rule ("savings") was calculated as (1 − the positivity rate) and represents the reduction in the fraction of patients needing evaluation when using the prediction rule compared with the universal screening approach (100% evaluated).

Data management and statistical analyses were performed using Stata version 15 (Stata Corporation, College Station, TX), JMP Pro version 13 (SAS Institute Inc., Cary, NC), and Python version 3.6 (Python Software Foundation, Wilmington, DE). The University of Vermont Committees on Human Subjects determined that the study did not constitute human subjects research.

Results

Overall, the prevalence of unhealthy drinking was 26%. The 43,545 records were randomly assigned to training (n = 28,262), validation (n = 6474), and test (n = 8809) sets. There were no significant differences among the 3 sets for any of the 38 variables. A total of 6% of values were missing and 23% of records were missing at least 1 variable.

Table 1 shows demographic and laboratory values by the reference drinking status (unhealthy versus low risk). On average, respondents in the unhealthy drinking category consumed 4.1 drinks per drinking day. In contrast, low-risk adults (including abstainers) had 1.5 drinks per drinking day. Individuals with unhealthy drinking were more likely to be younger, male, and current cigarette smokers. Although the differences in many clinical and laboratory values were statistically significant, they were small and unlikely to be clinically important.

Table 2 shows a comparison of the AUCs of the various methods across the training, validation, and test sets and the performance parameters for each model in the validation set. The random forest model produced the largest AUC in both the training set (0.85) and the validation set (0.80) and outperformed the other machine learning methods in sensitivity, specificity, PPV, NPV, overall accuracy, and savings in the validation set (see Figure 1). The random forest model was used to build the final clinical prediction rule. It was the only method used in the final test set.

Table 2. Performance of the Various Machine Learning Models in the Validation Set Using All 38 Variables*

Model	Training	Validation							Test						
	AUC (95% CI)	AUC (95% CI)	Sensitivity	Specificity	PPV	NPV	Overall Accuracy	Savings	AUC (95% CI)	Sensitivity	Specificity	PPV	NPV	Overall Accuracy	Savings
Universal Screening (No rule)	—	—	1.0	1.0	0.26	0.74	1.0	0%	—	1.0	1.0	0.26	0.74	1.0	0%
Random Forest	0.85 (0.84–0.86)	0.80 (0.79–0.81)	0.45	0.90	0.58	0.82	0.79	85%	0.78 (0.77–0.79)	0.50	0.88	0.55	0.83	0.76	75%
Support Vector Machines	0.81 (0.80–0.82)	0.77 (0.76–0.78)	0.34	0.89	0.50	0.79	0.74	82%	—	—	—	—	—	—	—
Neural Networks	0.79 (0.78–0.80)	0.78 (0.77–0.78)	0.36	0.90	0.58	0.80	0.76	82%	—	—	—	—	—	—	—
K-nearest Neighbors	0.78 (0.78–0.79)	0.75 (0.74–0.76)	0.35	0.84	0.45	0.78	0.71	79%	—	—	—	—	—	—	—
Decision Trees	0.77 (0.76–0.78)	0.75 (0.73–0.76)	0.34	0.90	0.56	0.79	0.75	83%	—	—	—	—	—	—	—
Logistic Regression	0.76 (0.75–0.77)	0.71 (0.70–0.73)	0.48	0.85	0.55	0.81	0.74	76%	—	—	—	—	—	—	—

*Sensitivity, specificity, PPV, NPV, Overall Accuracy, and Savings are all calculated at the selected optimum operating point in each case.
PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; CI, confidence interval.

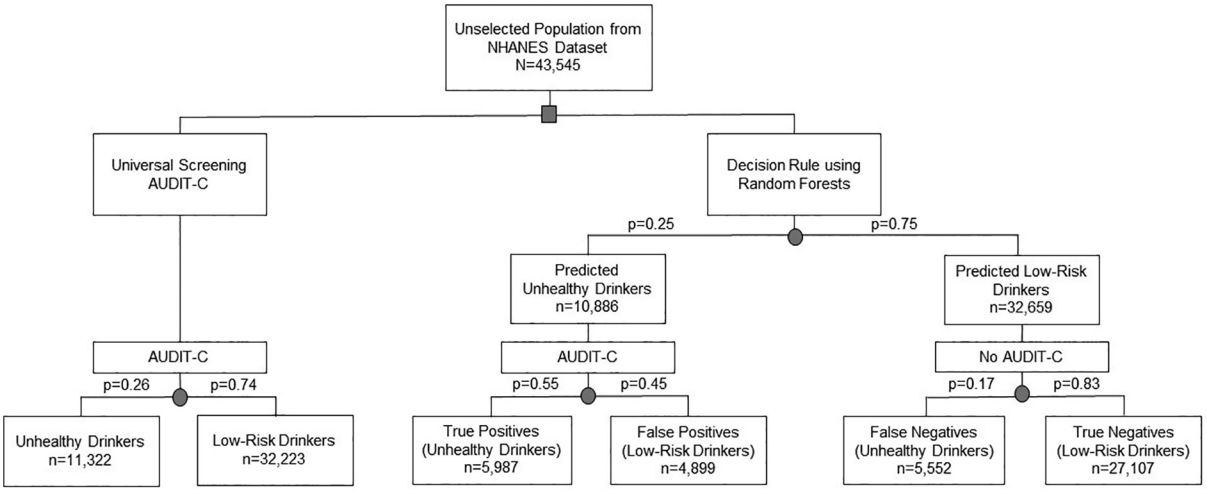
Figure 1. Random Forest AUC for training, validation, and test sets. Abbreviations: AUC, area under the receiver operating characteristic curve.

After selecting random forest as the final method, variables that contributed an information gain of <2% were dropped to create the most parsimonious model, ultimately including only 15/38 variables. The final model included the following predictors: age, current smoker, hemoglobin, sex, high-density lipoprotein, hematocrit, γ-glutamyl transpeptidase, mean cellular hemoglobin, uric acid, albumin, lactate dehydrogenase, mean corpuscular volume, systolic blood pressure, creatinine, and blood urea nitrogen (Table 3).

Table 3. Information Gain of Variables Used in the Final Prediction Model

Reference Group	Information Gain (%)
Age	28.1
Current smoker	10.7
Hemoglobin	7.7
Sex	7.3
High density lipoprotein	6.3
Hematocrit	6.0
Gamma-glutamyl transpeptidase	5.4
Mean cellular hemoglobin	4.8
Uric acid	4.4
Albumin	3.7
Lactate dehydrogenase	3.2
Mean corpuscular volume	3.2
Systolic blood pressure	3.1
Creatinine	3.1
Blood urea nitrogen	3.0

Figure 2. Population effect of using the clinical prediction rule to identify unhealthy drinking compared with universal screening. Abbreviations: NHANES, National Health and Nutrition Examination Survey; AUDIT-C, Alcohol Use Disorders Identification Test – alcohol consumption questions.



Compared with the presumed effects of universal screening (all patients are screened and all instances of unhealthy drinking are identified), the clinical prediction rule finds fewer unhealthy drinkers but at a much lower cost (see Figure 2). At a prevalence of 26% and at the optimum operating point, the clinical prediction rule has a sensitivity of 0.50, requiring that only 25% of the population undergo further evaluation (see Table 2). The PPV of 0.55 indicates that 55% of them are identified as having unhealthy drinking, compared with 26% of all patients identified with universal screening. By eliminating 75% of the population with a relatively low risk of unhealthy drinking, the model increases the prevalence of unhealthy drinking in the identified group and lowers the number assessed from 43,345 to 10,886 in this population.

With the same prediction rule, the operating point could be shifted along the receiver operating characteristic curve to prioritize sensitivity. For example, an alternate operating point prioritizing sensitivity could produce a sensitivity of 0.88, specificity of 0.49, PPV of 0.38, and NPV of 0.92. However, 61% of the population (n = 26,562) would need to be evaluated.

Discussion

We used commonly available laboratory, clinical, and demographic information from a nationally representative dataset to build a clinical prediction rule for unhealthy drinking. The analysis, which includes over 45,000 records, indicates that an automated tool can accurately identify unhealthy drinking by using commonly available secondary data, even with many missing values. Using a random forest model, we were able to predict unhealthy drinking with high specificity and modest sensitivity. Changing the operating point could allow for high sensitivity and modest specificity, if that were preferred. Random forest outperformed logistic regression and the other machine learning methods.

Prior studies on predicting unhealthy drinking have used classic statistical techniques with small data sets and limited computing power[23,25–27,46] compared with more modern methods. These prospective studies had control over the recruitment process and the ability to minimize missing data, which may have helped their prediction results. In contrast, the current study used a large existing dataset and analytical methods that accounted for missing data.

In the curated NHANES dataset, individual values were missing less than 5% of the time, but in EHRs, we would expect many more missing values. Some machine learning methods, especially random forest, consider and use missing data to create the most robust model.[47] Because all clinical data sources, including EHRs, have gaps, it is important that clinical prediction rules can account for missing data.

We tested logistic regression and multiple machine learning methods on the training and validation sets. Random forest outperformed all other methods, likely because it is particularly robust to outliers, missing data, and nonlinear relationships.[41] Although logistic regression is widely used in binary classification problems,[48] results in the medical literature are inconclusive about whether logistic regression can predict as well as machine learning methods.[28,29] A recent systematic review by Christodoulou et al[49] found no performance benefit of machine learning methods over logistic regression. However, logistic regression, and other methods that cannot handle missing data, are not practical in a clinical setting because users would either need to impute the missing data before applying the rule or abandon prediction for many cases. In the NHANES data, a particularly well-groomed dataset, only 77% of records had complete data. The choice of model for medical domains should be selected based on the problem to be solved; the understanding of the underlying biological, psychological, and social mechanisms; and the data available, rather than just whether the domain is medical or not.

The predictors of unhealthy drinking in the final model are biologically plausible and supported by the literature. Age, sex, smoking, and unhealthy drinking have been shown to be strongly correlated.[1,50] Alcohol use is associated with increased levels of high-density lipoprotein, reportedly through an increased transport rate of apolipoproteins A-I and A-II.[34] Others have used mean corpuscular volume, hemoglobin, γ-glutamyl transpeptidase, albumin, and systolic blood pressure in prediction models for heavy drinking.[23–25] Despite race and ethnicity being associated with alcohol use, they were removed a priori due to common misclassification problems, especially in EHR data.[51] To create the most parsimonious model, the random forest algorithm removed potential predictors that have a minimal effect on performance.

Universal screening results in many low-risk patients being offered an unnecessary intervention that PCPs are already reluctant to provide,[16,18,19,33] This clinical prediction rule prioritizes specificity over sensitivity and identifies patients who are likely to truly be drinking at an unhealthy level. Therefore, the population appropriate for follow-up assessment is greatly reduced compared with universal screening, freeing up time and resources. The trade-off is that some patients with unhealthy drinking are incorrectly categorized as low risk, missing an opportunity to intervene. If the setting warrants, the model can operate at a higher sensitivity, with correspondingly lower specificity.

This study has limitations. First, the NHANES sample is meant to be representative of the general population of adults in the United States, which may be different from those seeking primary care. The study population undoubtedly included some adults who would not be subjects for screening because, for example, they had a previously diagnosed alcohol use disorder. Second, the NHANES data may not be representative of EHR data, which would be used in practice. EHRs are likely to have much more missing data. However, random forest models are robust to missing data. Third, NHANES questionnaires were administered in person, possibly introducing social desirability response bias.[52] Therefore, alcohol and tobacco use may be underreported compared with self-report articles or electronic questionnaires. Because smoking was an important predictor in the model and alcohol use is the outcome, inaccurate reporting could result in misclassification. Nonetheless, self-report is the typical method for assessing smoking status and alcohol use in health care settings. Fourth, the prediction rule is not very transparent. Notably, it offers no single estimate of the relationship between any predictor and the outcome analogous to the odds ratio from a regression. A single predictor may seem to be harmful in some subgroups of patients and protective in others. Finally, we believe that this analysis overestimates the performance of universal screening because it assumes that all patients would be screened. In fact, a relatively low fraction of primary care patients are routinely screened with a validated tool such as the AUDIT.[17]

Conclusions

Motivated by critical barriers facing PCPs in identifying unhealthy drinking, we describe an alternative approach to routine universal screening: a clinical prediction rule based on existing data. This method could reduce the burden on PCPs and allow them to focus their attention on those who need it most. The virtue of the clinical prediction rule is not that it is perfectly accurate but that it is fast, inexpensive, unobtrusive, and identifies a subset of patients at a higher risk of unhealthy drinking.

To see this article online, please go to: http://jabfm.org/content/33/3/397.full.

References

1. Substance Abuse and Mental Health Services Administration (SAMHSA). 2015 National Survey on Drug Use and Health (NSDUH). Table 2.46B—alcohol use, binge alcohol use, and heavy alcohol use in past month among persons aged 12 or older, by demographic characteristics: percentages, 2014 and 2015. Available from: https://www.samhsa.gov/data/sites/default/files/NSDUH-DetTabs-2015/NSDUH-DetTabs-2015/NSDUH-DetTabs-2015.htm#tab2-46b. Published 2015. Accessed December 18, 2018.
2. National Institute on Alcohol Abuse and Alcoholism. Drinking levels defined. Available from: https://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/moderate-binge-drinking. Published 2015. Accessed December 18, 2018.
3. Abdul-Rahman AK, Card TR, Grainge MJ, Fleming KM. All-cause and cause-specific mortality rates of patients treated for alcohol use disorders: a meta-analysis. Subst Abus 2018;39:509–517.
4. Laramee P, Leonard S, Buchanan-Hughes A, Warnakula S, Daeppen JB, Rehm J. Risk of all-cause mortality in alcohol-dependent individuals: a systematic literature review and meta-analysis. EBioMedicine 2015;2:1394–404.
5. Holman CD, English DR, Milne E, Winter MG. Meta-analysis of alcohol and all-cause mortality: a validation of NHMRC recommendations. Med J Aust 1996;164:141–5.
6. Connor J. Alcohol consumption as a cause of cancer. Addiction 2017;112:222–8.
7. Dawson DA, Grant BF, Stinson FS, Chou PS. Psychopathology associated with drinking and alcohol use disorders in the college and general adult populations. Drug Alcohol Depend 2005;77:139–50.
8. Rehm J, Allamani A, Elekes Z, et al. Alcohol dependence and treatment utilization in Europe—a representative cross-sectional study in primary care. BMC Fam Pract 2015;16:90.
9. Rehm J, Room R, Graham K, Monteiro M, Gmel G, Sempos CT. The relationship of average volume of alcohol consumption and patterns of drinking to burden of disease: an overview. Addiction 2003;98:1209–28.
10. Huang C, Zhan J, Liu YJ, Li DJ, Wang SQ, He QQ. Association between alcohol consumption and risk of cardiovascular disease and all-cause mortality in patients with hypertension: a meta-analysis of prospective cohort studies. Mayo Clin Proc 2014;89:1201–10.
11. Sacks JJ, Gonzales KR, Bouchery EE, Tomedi LE, Brewer RD. 2010 national and state costs of excessive alcohol consumption. Am J Prev Med 2015;49:e73–e79.
12. Centers for Disease Control and Prevention (CDC). Alcohol and public health: Alcohol-Related Disease Impact (ARDI). Average for United States 2006–2010 alcohol-attributable deaths due to excessive alcohol use. Available from: https://nccd.cdc.gov/DPH_ARDI/Default/Report.aspx?T=AAM&P=f6d7eda7-036e-4553-9968-9b17ffad620e&R=d7a9b303-48e9-4440-bf47-070a4827e1fd&M=8E1C5233-5640-4EE8-9247-1ECA7DA325B9&F=&D=. Accessed December 1, 2018.
13. Moyer VA; Preventive Services Task Force. Screening and behavioral counseling interventions in primary care to reduce alcohol misuse: U.S. preventive services task force recommendation statement. Ann Intern Med 2013;159:210–8.
14. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Arch Intern Med 1998;158:1789–95.
15. Bradley KA, DeBenedetti AF, Volk RJ, Williams EC, Frank D, Kivlahan DR. AUDIT-C as a brief screen for alcohol misuse in primary care. Alcoholism Clin Exp Res 2007;31:1208–17.
16. Smith PC, Schmidt SM, Allensworth-Davies D, Saitz R. A single-question screening test for drug use in primary care. Arch Intern Med 2010;170:1155–60.
17. Denny CH, Hungerford DW, McKnight-Eily LR, et al. Self-reported prevalence of alcohol screening among U.S. adults. Am J Prev Med 2016;50:380–3.
18. Johnson M, Jackson R, Guillaume L, Meier P, Goyder E. Barriers and facilitators to implementing screening and brief intervention for alcohol misuse: a systematic review of qualitative evidence. J Public Health (Oxf) 2011;33:412–21.
19. Fortney J, Mukherjee S, Curran G, Fortney S, Han X, Booth BM. Factors associated with perceived stigma for alcohol use and treatment among at-risk drinkers. J Behav Health Serv Res 2004;31:418–29.
20. Beich A, Gannik D, Malterud K. Screening and brief intervention for excessive alcohol use: qualitative interview study of the experiences of general practitioners. BMJ 2002;325:870.
21. Onders R, Spillane J, Reilley B, Leston J. Use of electronic clinical reminders to increase preventive screenings in a primary care setting: blueprint from a successful process in Kodiak, Alaska. J Prim Care Community Health 2014;5:50–54.
22. Williams EC, Achtmeyer CE, Thomas RM, et al. Factors underlying quality problems with alcohol screening prompted by a clinical reminder in primary care: a multi-site qualitative study. J Gen Intern Med 2015;30:1125–1132.
23. Hartz AJ, Guse C, Kajdacsy-Balla A. Identification of heavy drinkers using a combination of laboratory tests. J Clin Epidemiol 1997;50:1357–1368.

24. Lichtenstein MJ, Burger MC, Yarned JWG, Elwood PC, Sweetnam PM. Derivation and validation of a prediction rule for identifying heavy consumers of alcohol. Alcoholism Clin Exp Res 1989;13:626–630.
25. Harasymiw J, Seaberg J, Bean P. Detection of alcohol misuse using a routine test panel: the early detection of alcohol consumption (EDAC) test. Alcohol Alcohol 2004;39:329–335.
26. Harasymiw JW, Vinson DC, Bean P. The early detection of alcohol consumption (EDAC) score in the identification of heavy and at-risk drinkers from routine blood tests. J Addict Dis 2000;19:43–59.
27. Korzec S, Korzec A, Conigrave K, Gisolf J, Tabakoff B. Validation of the Bayesian Alcoholism Test compared to single biomarkers in detecting harmful drinking. Alcohol Alcohol 2009;44:398–402.
28. Easton JF, Stephens CR, Angelova M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Comput Biol Med 2014;54:199–210.
29. Faisal M, Scally A, Howes R, Beatson K, Richardson D, Mohammed MA. A comparison of logistic regression models with alternative machine learning methods to predict the risk of in-hospital mortality in emergency medical admissions via external validation. Health Informatics J 2018; 1460458218813600.
30. Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. J Vasc Surg 2016;64:1515–1522, e1513.
31. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017;12:e0174708.
32. Arostegui I, Gonzalez N, Fernández-de-Larrea N, et al. Combining statistical techniques to predict postsurgical risk of 1-year mortality for patients with colon cancer. Clin Epidemiol 2018;10:235–251.
33. D'Amico EJ, Paddock SM, Burnam A, Kung FY. Identification of and guidance for problem drinking by general medical providers: results from a national survey. Med Care 2005;43:229–236.
34. De Oliveira E, Foster D, McGee Harper M, et al. Alcohol consumption raises HDL cholesterol levels by increasing the transport rate of apolipoproteins A-I and A-II. Circulation 2000;102:2347–2352.
35. Worrall S, Jersey J, Wilce PA, Seppä K, Hurme L, Sillanaukee P. Relationship between alcohol intake and immunoglobulin a immunoreactivity with acetaldehyde-modified bovine serum albumin.Alcoholism Clin Exp Res 1996;20:836–840.
36. Steyerberg EW. Clinical prediction models: a practical approach to development, validation and updating. New York (NY): Springer; 2010.
37. Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 1998;2:121–167.
38. Zhang GP. Neural networks for classification: a survey. IEEE Trans Syst Man Cy C 2000;30:451–462.
39. Altman NS. An Introduction to Kernel and nearest-neighbor nonparametric regression. Am Stat 1992; 46:175–185.
40. Quinlan JR. Induction of decision trees. Mach Learn 1986;1:81–106.
41. Breiman L. Random forests. Mach Learn 2001; 45:5–32.
42. Friedman J, Tibshirani R, Hastie T. The elements of statistical learning. New York (NY): Springer New York Inc.; 2001.
43. Zhang Y, Xin Y, Li Q, et al. Empirical study of seven data mining algorithms on different characteristics of datasets for biomedical classification applications. Biomed Eng Online 2017;16:125.
44. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
45. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. Burlington (MA): Morgan Kaufmann Publishers Inc.; 2011.
46. Korzec A, de Bruijn C, van Lambalgen M. The Bayesian Alcoholism Test had better diagnostic properties for confirming diagnosis of hazardous and harmful alcohol use. J Clin Epidemiol 2005;58:1024–1032.
47. Gaudard M, Ramsey P, Stephens M. Interactive data mining and design of experiments: The JMP partition and custom design platforms. Cary, North Carolina: North Haven Group, LLC; 2006. http://islab.soe.uoguelph.ca/sareibi/PROJECTS_dr/GRAD_FUTURE_dr/docs/Interactive_DataMining.pdf.
48. Held E, Cape J, Tintle N. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. BMC Proc 2016;10:141–145.
49. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22.
50. Istvan J, Matarazzo JD. Tobacco, alcohol, and caffeine use: a review of their interrelationships. Psychol Bull 1984;95:301–326.
51. Klinger EV, Carlini SV, Gonzalez I, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med 2015;30:719–723.
52. Davis CG, Thake J, Vilhena N. Social desirability biases in self-reported alcohol consumption and harms. Addict Behav 2010;35:302–311.

Appendix. Selected Machine Learning Methods for Classification of Unknown Cases into Mutually Exclusive Categories

Method	Advantage	Disadvantage
Random forest	• Low computational cost • Uses missing data to inform model • Can handle large number of records and variables • Provides estimates of the information gained by each input variable • Works well with nonlinear data	• Not ideal for rare outcomes • Very difficult to interpret individual variable contributions to classification • Time consuming hyperparameter tuning • Overfitting of data
Support Vector Machines	• Low computationally cost • Effective when number of variables> number of records (very wide data)	• Need a clear margin of separation between outcomes (unhealthy drinking *vs* low-risk) • Time consuming hyperparameter tuning • Not efficient with large number of records
Neural Networks	• Works well with nonlinear data • Extremely useful with large number of predictors (high dimensionality (e.g. image data)) • Any numeric data can be used	• High computational cost during training • Time consuming hyperparameter tuning • Need relatively large number of records for training set • Very difficult to interpret individual variable contributions to classification • Must have many records per variable • Overfitting of data
K-nearest neighbors	• Very simple construction requiring minimal specifications (a.k.a. hyperparameters) • Intuitive methodology	• High computational cost • Challenging with large number of variables (wide data) • Cannot handle imbalanced data • Very sensitive to outliers • Cannot handle missing data
Decision Trees	• Can handle missing data • No data preprocessing needed • Provides highly intuitive explanation over the prediction	• Highly biased to training set • Relatively inaccurate compared to other models
Logistic Regression	• Common and understood by most • Relatively easy to implement • Loss function is always convex	• Proper selection of features is required • Cannot handle missing data • Needs data preprocessing and handling to cover non-linear data • Cannot handle large number of categorical predictors