Automated Identification of Unhealthy Drinking Using Routinely Collected Data: A Machine Learning Approach

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

INTRODUCTION

- Unhealthy drinking is highly prevalent and can lead to serious health and social consequences.
- Identifying unhealthy drinkers can be time-consuming for primary care providers.
- Unhealthy drinking is under-identified and under-treated.
- An automated tool for identification would allow attention to be focused on patients most likely to need care and therefore increase efficiency and effectiveness.

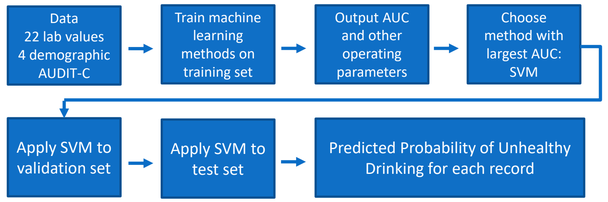
STUDY OBJECTIVE

To build a clinical prediction tool for unhealthy drinking based solely on routinely collected demographic and laboratory data.

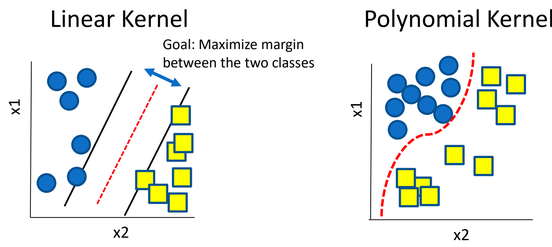
METHODS

- Demographic and laboratory data on 13,833 adults seen at the University of Vermont Medical Center, 2011-2017.
- Predictors
  - Demographic information
    - Smoking status, gender, age, race/ethnicity
  - 22 Lab values from routine clinical chemistry and hemograms.
- Reference test
  - Unhealthy drinking measured by AUDIT-C
- Definitions
  - > 3 drinks per day or > 7 drinks per week for women
  - > 4 drinks per day or > 14 drinks per week for men
- 22 lab values, 4 demographic variables.
- Training set (68%), validation set (12%), test set (20%).
- Logistic regression, support vector machines (SVM), k-nearest neighbor, and random forests were used to build clinical prediction models.
- Model with largest area under the curve (AUC) was selected.
- Support vector machine with polynomials of degree 3 produced the largest AUC.
- An operating point with greater specificity was prioritized.

BUILDING THE DECISION RULE

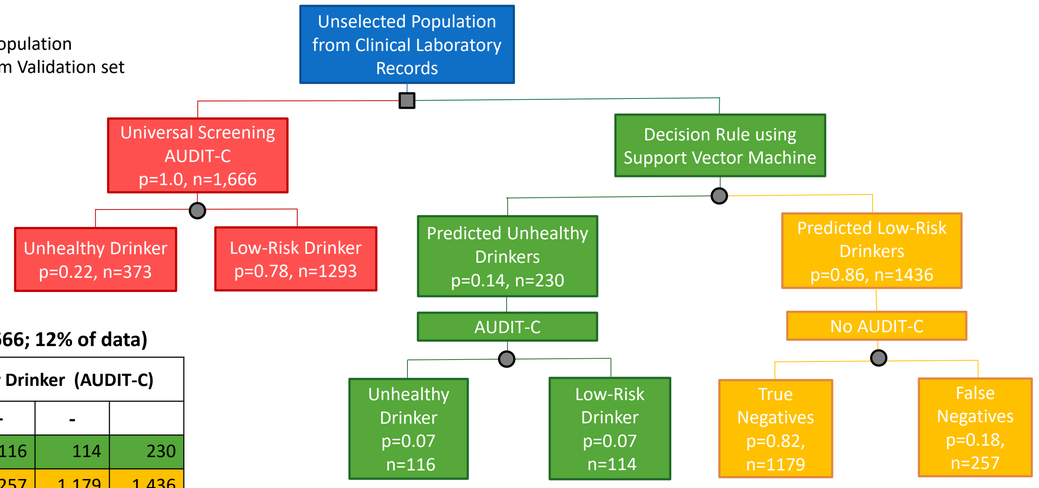


SUPPORT VECTOR MACHINES



POPULATION EFFECT OF USING DECISION RULE TO IDENTIFY UNHEALTHY DRINKING

Key
p : Proportion of total population
n : Relevant number from Validation set
■: Policy choice
●: Probability



Validation Set (n=1,666; 12% of data)

	Unhealthy Drinker (AUDIT-C)		
Predicted Unhealthy Drinker (SVM)	+	-	
+	116	114	230
-	257	1,179	1,436
	373	1,293	1,666

***Decision rule decreases number needed to screen from 1,666 to 230**

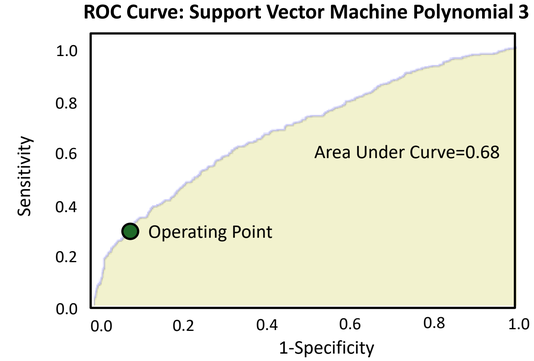
RESULTS

Cohort Characteristics (n=13,883; 100% of data)

		n (%)
Sex (female)		6,778 (48.8)
Race/Ethnicity	White	13,107 (94.4)
	Declined	348 (2.5)
	Other	428 (3.1)
Smoking Status (yes)		1,481 (10.8)
Prevalence of Unhealthy Drinking		3,102 (22.3)
*Mean Age ± SD		59.5 ± 15.1
*Mean HDL Cholesterol ± SD		57.5 ± 19.6
*Mean White Blood Count ± SD		7.5 ± 3.4
*Mean Hemoglobin ± SD		13.5 ± 1.8

*Influential predictors

ROC Curve: Support Vector Machine Polynomial 3



RESULTS

- We chose an optimum operating point that produces:
  - sensitivity 31.1%
  - specificity 91.2%
  - positive predictive value 50.4%
  - negative predictive value 82.1%
  - Change in prevalence of unhealthy drinking from 22% to 50%
  - Reduced the target population by 86%
- The most influential predictors were age, HDL cholesterol, white blood count, and hemoglobin.

DISCUSSION

- A clinical prediction model using machine learning methods can improve case identification.
- These results have the potential to shift the case identification paradigm from passive/universal screening to a data-driven approach.

CONCLUSION

The virtue of the clinical prediction rule is not that it is perfectly accurate but that it is fast, inexpensive, unobtrusive, and identifies subjects with a higher prevalence of unhealthy drinking than the normal population.

LIMITATIONS

- Primary care providers were not practicing universal screening.
  - Prevalence of unhealthy drinking 60% among those with AUDIT-C
- Sample may not be generalizable
- Sensitivity and specificity of AUDIT-C

ACKNOWLEDGEMENTS & CONTACT

Thanks to the many patients whose data we used and to the NIH for funding.
SRRT 1R41AA025297-01A1

Gail Rose, PhD <gail.rose@uvmhealth.org>
Levi Bonnell, MPH <levi.bonnell@med.uvm.edu>
Given S-459
University of Vermont
89 Beaumont Ave, Burlington, VT, 05404