

Accepted Manuscript

Co-Occurring Risk Factors for Current Cigarette Smoking in a U.S. Nationally Representative Sample

Stephen T. Higgins, Allison N. Kurti, Ryan Redner, Thomas J. White, Diana R. Keith, Diann E. Gaalema, Brian L. Sprague, Cassandra A. Stanton, Megan E. Roberts, Nathan J. Doogan, Jeff S. Priest

PII: S0091-7435(16)30004-4
DOI: doi: [10.1016/j.ypmed.2016.02.025](https://doi.org/10.1016/j.ypmed.2016.02.025)
Reference: YPMED 4548

To appear in: *Preventive Medicine*

Received date: 11 December 2015
Revised date: 16 February 2016
Accepted date: 18 February 2016

Please cite this article as: Higgins Stephen T., Kurti Allison N., Redner Ryan, White Thomas J., Keith Diana R., Gaalema Diann E., Sprague Brian L., Stanton Cassandra A., Roberts Megan E., Doogan Nathan J., Priest Jeff S., Co-Occurring Risk Factors for Current Cigarette Smoking in a U.S. Nationally Representative Sample, *Preventive Medicine* (2016), doi: [10.1016/j.ypmed.2016.02.025](https://doi.org/10.1016/j.ypmed.2016.02.025)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



Co-Occurring Risk Factors for Current Cigarette Smoking in a U.S. Nationally Representative**Sample**

Stephen T. Higgins, PhD¹, Allison N. Kurti, PhD¹, Ryan Redner, PhD¹, Thomas J. White, PhD¹, Diana

R. Keith, PhD¹, Diann E. Gaalema, PhD¹, Brian L. Sprague, PhD¹, Cassandra A. Stanton, PhD²

Megan E. Roberts, PhD³, Nathan J. Doogan, PhD³, Jeff S. Priest, PhD¹

Vermont Center on Tobacco Regulatory Science, University of Vermont¹, Westat², and Center of

Excellence in Regulatory Tobacco Science, The Ohio State University³

Word count:

Abstract: 250

Text: 3,689

Correspondent:

Stephen T. Higgins, PhD

email: Stephen.Higgins@uvm.edu

phone: 802-656-9615

Abstract

Introduction: Relatively little has been reported characterizing cumulative risk associated with co-occurring risk factors for cigarette smoking. The purpose of the present study was to address that knowledge gap in a U.S. nationally representative sample. **Methods:** Data were obtained from 114,426 adults (≥ 18 years) in the U.S. National Survey on Drug Use and Health (years 2011-13). Multiple logistic regression and classification and regression tree (CART) modeling were used to examine risk of current smoking associated with eight co-occurring risk factors (age, gender, race/ethnicity, educational attainment, poverty, drug abuse/dependence, alcohol abuse/dependence, mental illness). **Results:** Each of these eight risk factors was independently associated with significant increases in the odds of smoking when concurrently present in a multiple logistic regression model. Effects of risk-factor combinations were typically summative. Exceptions to that pattern were in the direction of less-than-summative effects when one of the combined risk factors was associated with generally high or low rates of smoking (e.g., drug abuse/dependence, age ≥ 65). CART modeling identified subpopulation risk profiles wherein smoking prevalence varied from a low of 11% to a high of 74% depending on particular risk factor combinations. Being a college graduate was the strongest independent predictor of smoking status, classifying 30% of the adult population. **Conclusions:** These results offer strong evidence that the effects associated with common risk factors for cigarette smoking are independent, cumulative, and generally summative. The results also offer potentially useful insights into national population risk profiles around which U.S. tobacco policies can be developed or refined.

Key terms: Risk factors, co-occurring risk factors, adults, U.S. nationally representative sample, cigarette smoking, current smokers, multiple logistic regression, classification and regression tree (CART), educational attainment

There have been substantial decreases in U.S. national smoking prevalence since the mid 1960's, but unfortunately these decreases have been unevenly distributed in the general population (Schroeder & Koh, 2014). Substantial reductions have been noted in some subpopulations (e.g., more affluent non-Hispanic Whites), but relatively little change has occurred in others (e.g., those with substance use disorders), and increases in still others (e.g., economically disadvantaged women) (Chilcoat, 2009; Fiore et al., 2008; Schroeder & Koh, 2014). These uneven changes in smoking prevalence underpin considerable current interest in understanding individual differences in risk for cigarette smoking and other forms of tobacco and nicotine use.

Monitoring the extent to which prevalence of smoking or use of other tobacco products differs by risk factors is now recognized to be an important element of tobacco control (Fiore et al., 2008) and regulatory science (Ashley et al., 2014). The overarching purpose of the present study is to begin characterizing the effects of co-occurring risk factors for cigarette smoking. We focus on cigarette smoking because it remains the most prevalent, toxic, and costly form of tobacco and nicotine use (U.S. DHHS, 2014). We know of no exhaustive set of risk factors for cigarette smoking, although gender, age, race/ethnicity, educational attainment, poverty status, substance use disorders, and mental illness are each well documented in the literature and are examined in the present study (Fiore et al., 2008; Higgins et al., 2015; Higgins & Chilcoat, 2009; Schroeder & Koh, 2014). While each of these risk factors inevitably co-occurs with some arrangement of the others (i.e., gender always co-occurs with chronological age, educational attainment, race/ethnicity, etc.), there has been relatively little research reported explicitly characterizing the combined effects of co-occurring risk factors for cigarette smoking. Knowing whether effects of co-occurring risk factors are independent of each other (i.e., summative), or whether some may offset risks associated with others (i.e., less-than-summative / antagonistic), or perhaps increase risk in a multiplicative (i.e., synergistic) manner is important to the development of evidence-based tobacco policy and is the overarching purpose of the present study.

Also of interest is empirically examining the relative strength of these common risk factors and how cumulative risk varies across particular risk factor profiles.

In a prior literature review on gender differences in prevalence of cigarette smoking and use of other nicotine and delivery products in the U.S., gender differences in risk were noted to generally act independently of the influence of other co-occurring risk factors, that is, gender and the other risk factors appeared to act in a cumulative and summative manner (Higgins et al., 2015). However, these were qualitative observations regarding patterns in previously published articles, none of which was explicitly designed to characterize the combined effects of co-occurring risk factors.

The present study was designed to build upon the initial findings described above by examining (a) a broader range of co-occurring risk factors than those involving gender, (b) statistically examining the independent and combined effects of common risk factors for smoking, and (c) identifying particularly low- and high-risk profiles. We know of no prior studies specifically on this topic regarding risk for current cigarette smoking, although studies characterizing effects of co-occurring risk factors are common in other areas of health research (e.g., Park et al., 2009; Schnohr et al., 2002). The present study was conducted using what at the time of study initiation was the most recent three years (2011-2013) of the National Survey on Drug Use and Health (NSDUH) (SAMHSA, 2012; 2013; 2014), a cross-sectional survey that has been used effectively in prior studies examining prevalence of cigarette smoking across various socio-demographic and psychiatric risk factors (e.g., Gfroerer et al., 2013; Redner et al., 2014a; Redner et al., 2014b; White et al., 2015).

Methods

Data Source

The NSDUH is a nationally representative survey of the U.S. non-institutionalized population aged ≥ 12 years that measures prevalence and correlates of drug use (Center for Behavioral Health Statistics and Quality, 2014). Detailed descriptions of survey procedures have been provided for each of the survey years (SAMHSA, 2012, 2013, 2014). Only individuals aged ≥ 18 years were included in the

present study so that all participants were of legal age to purchase cigarettes. Across each of the survey years, NSDUH recruitment was completed using a multistage area probability sample design in which a predetermined number of participants were randomly recruited by address within each state.

Respondents completed computer- and audio-assisted structured interviews. Respondents were selected from the civilian non-institutionalized population, including group homes, shelters, and college dormitories. Individuals on active military duty, in residential drug treatment programs, in jail, or homeless without residence were excluded. The present study included 114,426 adult respondents interviewed during 2011 ($N = 39,133$), 2012 ($N = 37,869$), and 2013 ($N = 37,424$). Annual weighted response rates were 74.4%, 73.0%, and 71.7% for the 2011, 2012, and 2013 surveys, respectively. Data were weighted during analysis to adjust for the differential probability of both selection and response.

Current smoker status was defined as smoking all or part of a cigarette within the 30 days preceding the interview and ≥ 100 cigarettes lifetime. The six racial/ethnic categories used in the survey were mutually exclusive. Persons identified as Hispanic might be of any race while persons identified as White, Black, Asian, American Indian/Alaska Native, or Other were all non-Hispanic. The category “Other” included Native Hawaiians or Other Pacific Islanders and individuals endorsing two or more races. Poverty status (living below or at/above the federal poverty line) was defined using poverty thresholds published by the U.S. Census Bureau. Any mental illness was defined as having a mental, behavioral, or emotional disorder in the past 12 months, excluding developmental or substance use disorders. To assess the presence of any mental illness, respondents aged ≥ 18 years answered a series of 14 questions that made up two scales measuring psychological distress (Kessler-6) and disability (World Health Organization Disability Assessment Schedule). Scores on these two scales were used to determine any mental illness status based on a statistical model developed from clinical interviews that assessed disorders based on criteria in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV). Past year alcohol and illicit drug abuse/dependence diagnoses were also based on DSM-IV criteria.

Statistical Methods

Sample-adjusted frequencies and confidence intervals (CIs) were generated across all respondents ≥ 18 years of age. Variables of interest were determined based on previously identified demographic and socioeconomic predictors of smoking, including age, sex, education, race/ethnicity, poverty status, past year mental illness, alcohol abuse/dependence, and illicit drug abuse/dependence.

Associations of risk factors with current smoking status were examined in separate analyses. For each risk factor, weighted, univariate logistic regression was used to determine which variables would be included in subsequent multivariable models and to conduct an initial comparison of odds ratios (with CIs) across variable levels predicting current smoker status. PROC SURVEYLOGISTIC in SAS (SAS Institute, Cary, NC) was used to conduct the analyses, relying on maximum likelihood estimation and the Fisher scoring algorithm. Variances were estimated using Taylor series linearization.

Multivariable logistic regression analyses predicting current smoker status were conducted using all variables from univariate logistic regression analyses that significantly predicted smoking status, initially without including any interaction terms. Again, odds ratios (with CIs) were generated. Analyses were repeated examining all possible two-way interactions across the eight risk factors (maximum of 28 risk-factor combinations each across analyses predicting smoking status). Interactions were first tested individually. Interactions that contributed significantly to predicting current smoker status were added to the multivariable model. Interactions that remained significant predictors were retained in the final regression model. These analyses were conducted using SAS 9.4 software (SAS Institute, Cary, NC). Across all tests, statistical significance was defined as $p < .001$ (2-tailed) to correct for multiple comparisons corresponding to investigating interactions.

Interaction results from the final regression model were used to initially characterize the combined effects of the risk factors. Where an interaction term was non-significant between two risk factors that each produced significant main effects, combined effects were categorized as summative (i.e., sum of the independent effects of the respective risk factors). Where significant interactions were

noted between predictors, they were categorized as representing a deviation from summation. Further determination of whether an interaction represented a less-than- or greater-than-summative effect was determined by reviewing graphic displays of the relevant odds ratios (see Figure 2).

Lastly, a classification and regression tree (CART) analysis (Breiman et al., 1984) was conducted to supplement multiple logistic regression modeling, using the same eight independent variables to predict current smoking. CART analysis is a nonparametric procedure for dividing a population of interest into mutually exclusive subgroups based on a dependent variable of interest such as current smoking status in the present study (Lemon et al., 2003) and, in the process, identifying independent variables with the most explanatory power in accounting for that dependent variable. The process begins by identifying the single most important independent variable for dividing the total sample (parent node) into two groups (child nodes), using a predetermined branching criterion. Nodes are split based on their purity using the Gini impurity function (Breiman et al., 1984). A “pure” node has no variability in the dependent variable. A completely “impure” node has a conditional probability of $p(k|t) = 0.5$, where k refers to the dependent variable and t refers to the node (Lei et al., 2015). A splitting or branching criterion “selects the split that has the largest difference between the impurity of the parent node and a weighted average of the impurity of the two child nodes” (Lemon et al., 2003, p. 174). Given the dependent variable was binary, we used the Gini impurity function to split nodes, repeating the process recursively with every subsample, until the subsample reached a minimum size or no further splits could be made. The tree was built using R’s rpart package (R Core Team, 2013; Therneau et al., 2013). We used the classification method in R, given the dependent variable was binary, and included survey weights, given the multi-stage sampling procedures of the NSDUH. A fully saturated tree was produced initially, and then pruned by selecting the complexity parameter that minimized cross-validation error and setting a minimum sample size in terminal nodes of $n = 1,000$.

Results

Logistic Regression Analyses

Overall prevalence of current smokers in this adult sample was 21.6% (Table 1, left-most column). Each of the eight risk factors significantly increased the odds of being a current smoker in univariate logistic regression (Table 1, center-most columns). Each of those risk factors also remained significant in a multivariate logistic regression model adjusting for the influence of the others, demonstrating significant independent associations with smoking status (Table 1, right-most columns). The largest increase in the adjusted odds of smoking was seen with educational attainment, followed by age, past year illicit drug abuse/dependence, race/ethnicity, past year alcohol abuse/dependence, income below federal poverty level, mental illness, and gender.

Results from including the two-way interaction term in the final multivariate logistic regression model were used for an initial characterization of combined effects of these eight significant risk factors (Table 2). Across all of the possible two-way interactions, 20 of 28 (71%) were not significant, meaning that the combined effects of these risk factors were summative (i.e., they did not deviate significantly from a summation of the effects of the two risk factors when considered alone) (Figure 1).

The eight (29%) significant interactions involved each of the eight risk factors, but mostly race/ethnicity which has 6 levels in the present study and was involved in five of the eight significant interactions (Table 2). Age was involved in three interactions and the six other risk factors were each involved in one interaction (Table 2).

As is illustrated in Figure 2, these interactions largely involved less-than-summative effects where one of the two risk factors was associated with generally high or low smoking prevalence. As shown in Panel A, for example, the typical pattern of Blacks having lower odds than Whites for being a current smoker was present in the 18-25 years but not the ≥ 65 years age bracket where smoking rates are generally low. Similarly, as shown in Panel B, the typical pattern of Hispanics having lower smoking rates than Whites was discernible among those with less than a high school education but not

among college graduates where smoking rates are generally low. In Panel C we see a similar example where the effects associated with having less than a high education are discernible among those in the 18-25 years but not the ≥ 65 years age bracket where smoking rates are generally low. As a final example, in Panel D we see increases in the odds of smoking associated with alcohol abuse/dependence among those without but not those with illicit drug abuse/dependence where smoking rates are generally high. We saw no evidence of interactions related to multiplicative effects of combining risk factors.

Use of logistic regression to characterize more than two-way interactions when dealing with 8 risk factors becomes impractical in terms of interpretation, and thus we used the CART analysis for that task and also for comparing the relative strength of the eight risk factors. Regression trees are adept at illustrating important risk factor combinations or profiles (Austin, 2007).

Classification and Regression Tree (CART) Analyses

The CART analysis identified educational attainment as the strongest risk factor followed by age, race/ethnicity, drug abuse/dependence, alcohol abuse/dependence, poverty level, mental illness, and gender. Figure 3 shows a pruned classification tree modeling changes in smoking prevalence associated with the various risk-factor combinations. The graphic is designed to represent an inverted tree.

The rectangle shown at the top of Figure 3 is referred to as the root node. It represents 100% of the U.S. adult non-institutionalized population as indicated in the bottom row of information displayed in the node, with 78% non-smokers and 22% smokers as indicated in top row of information displayed in the node. The 1st split of the entire population was based on whether someone was or was not a college graduate (dashed lines immediately below root node). College graduates branched leftward and downward to a terminal node (no further splitting/classification possible) with a smoking prevalence of only 11%, one half that seen in the overall population. Note that terminal nodes also display a 3rd piece of information not shown in the root or child nodes, which is the percentage of the current smoking population that is represented by that node (bottom-most row). This terminal node representing college graduates includes prevalence rates of 11% current smokers and 89% non-smokers, and represents 30%

of the U.S. adult non-institutionalized population and 15% of adult current smokers within that population. The 70% of the overall population (smokers & non-smokers) that had less than a college education branched rightward and downward to a child node (further splitting/classification possible) where prevalence of current smoking increased to 26% corresponding to removal of the relatively low-risk college graduates from the sample.

The 2nd branching was based on chronological age, dividing all those in the adult U.S. non-institutionalized population with less than a college education by whether they were ≥ 65 vs. 18-64 years of age. Those ≥ 65 years branched leftward and downward to a terminal node. Note that smoking prevalence in those ≥ 65 years is lowest among all age brackets. As such, the 11% smoking prevalence in this node was equal to that seen among college graduates despite this subgroup having lower educational attainment. Those whose age was below 65 years branched rightward and downward to a child node where smoking prevalence rates increased further to 30% corresponding to having removed those in the oldest age bracket.

The next branching of this subgroup with less than a college education and ages between 18 and 64 years was based on the absence vs. presence of past year drug abuse/dependence. Those without drug abuse/dependence branched leftward and downward to a child node where smoking prevalence decreased to 28% corresponding to having removed the relatively small subgroup with drug abuse/dependence. The subgroup with past year drug abuse/dependence moved rightward and downward to a child node where smoking prevalence increased to 67%, more than three-fold above the overall 22% smoking prevalence rate seen in the entire adult population. Further branching of this subgroup was based on additional age levels followed by race/ethnicity levels resulting in three separate terminal nodes. Note that all individuals represented in those three terminal nodes had less than a college education and past year drug abuse/dependence, but smoking prevalence nevertheless varied between 46% and 74% depending on the particular age and race/ethnicity levels with which those risk

factors were combined. Those terminal nodes each represented 1% or less of the entire population and thus a relatively small (1-3%) overall proportion of current smokers.

This same general branching process was repeated until the entire study population was classified in risk profiles across the 13 terminal nodes in the bottom row. The four left-most nodes represent ~ 90% of the U.S. adult non-institutionalized population and the 9 right-most nodes ~ 10%. Note that the majority of all current smokers (74%) are represented in the four left-most nodes even though smoking prevalence rates are much lower in those nodes. An alternative way of characterizing this distribution is that smoking is strikingly overrepresented in the 9 rightmost terminal nodes such that 26% of all current smokers are represented among only 10% of the total population.

Balancing smoking prevalence against proportion of the population represented, the terminal node or risk profile that represents the largest proportion of all current smokers is the fourth from left that comprises individuals with high school or some college education, ages 18-64 years, all race/ethnicity levels except Asian & Hispanic, and no past year alcohol or drug abuse/dependence. That node represents 34% of the entire population and 43% of all adult current smokers.

Discussion

The present study was conducted to follow up on observations reported as part of a literature review on gender differences where risk for cigarette smoking appeared to change in a cumulative and summative manner when gender was considered in combination with other co-occurring risk factors (Higgins et al., 2015). The present results confirm those earlier observations and extend them to additional risk-factor combinations beyond gender. Results from the multiple logistic regression analyses provide clear evidence that each of the eight risk factors examined act as independent predictors. Results from testing all possible two-way interactions among those eight independent predictors indicated that they usually did not interact significantly (i.e., 20 of 28 tests were non-significant). That is, these risk factors generally acted in a cumulative and summative manner where effects of the risk factors in combination did not change significantly from those observed when each

was examined as single predictors in the adjusted model. In those instances where there were significant interactions, the general pattern was that one of the combined risk factors was associated with diminished effects compared to when examined alone (less-than-summatve effects) (Figure 2). The CART analysis provided numerous opportunities to observe orderly upward and downward changes in smoking prevalence across population subgroups corresponding to changes in risk-factor combinations. Considered together, these results provide an empirical framework for understanding the striking differences in smoking prevalence observed across population subgroups and for making predictions about the likely effects of novel risk-factor combinations.

Three more specific points about the risk profiles from the CART analysis merit comment. First, there was only one instance where a single risk factor--actually a single risk-factor level-- acted as a stand-alone risk profile. That was being a college graduate. Moreover, this single risk-factor level classified 30% of the U.S. adult population demonstrating considerable reach. Educational attainment was also associated with the largest changes in the odds of smoking in the multiple logistic regression. These results underscoring the importance of a college education as a protective factor and identifying educational attainment more generally as the strongest risk factor among the eight examined is consistent with the emphasis that has been previously placed on educational attainment in discussing prevention of smoking among youth and young adults (e.g., see Chapter 2, U.S., DHHS, 2012) as well as addressing smoking risk among women and other vulnerable populations (Chilcoat, 2009; Graham et al., 2007; Higgins et al., 2009; Hiscock et al., 2012; Kandel et al., 2009). Also important to underscore is that educational attainment is a modifiable risk factor. Considering the pervasive and robust associations between educational attainment and smoking risk, there are grounds for a broader approach to tobacco control policy that encompasses more distal risk factors like general educational attainment in addition to the more proximal and conventional tobacco-control foci (Graham et al., 2007; Higgins et al., 2009; Kandel et al., 2009).

Second, the risk profile corresponding to the 4th terminal node in Figure 3 (counting from left to right) warrants further comment. That risk profile represents adults with a high school or some college education level, chronological age in the 18-64 years bracket, Black, White, American Indian/Native Alaskan, or Other race ethnicity, and no past year alcohol or drug abuse/dependence. By a factor of 2.9- to 43-fold, the profile included more current smokers (43%) than any of the 12 other risk profiles identified in the CART analysis. This profile represents a large segment of the U.S. adult population who educationally fall into the unskilled and skilled labor socioeconomic class and who for reasons that are not well understood have been relatively less responsive to efforts to reduce smoking (e.g., Asfar et al., 2015). Prior reports underscoring the relatively high smoking rates in blue-collar and service occupations have directed attention to the workplace setting as an important potential focal point for tobacco control and regulatory efforts that would have the potential to reach many within this subgroup (CDC, 2011). However, additional and complementary campaigns specifically targeting this risk-profile subgroup through other channels and contexts (social media, health care providers, community/civic organizations) are likely to be needed as well. This is a risk profile subgroup with which tobacco reduction efforts will need to improve considerably if the smoking-related goals of Healthy People 2020 are to be realized (Office of Disease Prevention and Health Promotion, 2015).

Third, the subgroups represented across the nine rightmost terminal nodes in the bottom row of Figure 3 merit further comment as well. Smoking prevalence rates ranged from 37% to 74% across those terminal nodes corresponding to a relatively larger number of risk factors and higher risk-factor levels than in the other nodes. Collectively, those nine nodes accounted for only 10% of the population but 26% of all current smokers. While not explicitly examined in this study, smokers with these profiles are more likely to be nicotine dependent and are at increased risk for adverse health impacts of smoking (Higgins & Chilcoat, 2009; Hiscock et al., 2012). The lower educational attainment levels and higher rates of alcohol and drug use disorders represented in those nodes also make it likely that other medical co-morbidities will be present further increasing vulnerability to the adverse health impacts of smoking

(Cutler & Lleras-Muney, 2010; Gaalema et al., 2015; Hser et al., 2001; Niaura et al., 2012; Rowa-Dewar et al., 2015; Schroeder, 2007). These are patterns that contribute directly to the unsettling problem of health disparities (Higgins, 2014; Schroeder, 2007). If considered only in terms of absolute numbers of smokers, those nine nodes may appear to warrant less attention or resources. However, when considered in terms of the overall potential morbidity, mortality, and healthcare cost impacts involved, those nodes represent risk profiles where more effective strategies for reducing smoking are sorely needed. For potential behavioral economic and pharmacological strategies for doing interested readers may want to see reports by Davis et al. (2016, this issue) and Tidey et al. (2016, this issue).

There are several limitations of the present study that should be acknowledged. First, the observational research design used in the present study cannot support causal inferences. Second, the NSDUH excludes several groups with relatively high smoking prevalence rates including individuals in the active military, jail, or homeless, which may limit generalizability of the results to those subgroups. Moreover, we excluded adolescents thereby potentially limiting generalizability to that important subgroup. Third, the NSDUH is a cross-sectional survey and thus does not permit examination of associations within individuals over time. Extending this research on co-occurring risk factors to a longitudinal survey such as the Population Assessment of Tobacco and Health (PATH, 2016) will be an important future research direction to assess whether cumulative risk is generally summative when examined prospectively over time. Fourth, the present study did not include tobacco products other than cigarettes, which will be an important gap to address in future research. Those limitations notwithstanding, we believe that the present study provides important new knowledge regarding effects associated with co-occurring risk factors for current cigarette smoking that has the potential to inform and advance evidence-based tobacco control and regulatory policy efforts.

Funding

This project was supported in part by Tobacco Centers of Regulatory Science (TCORS) award P50DA036114 from the National Institute on Drug Abuse (NIDA) and Food and Drug Administration (FDA), TCORS award P50CA180908 from the National Cancer Institute (NCI) and FDA, Center for Evaluation and Coordination of Training and Research award U54CA189222 from NCI and FDA, Institutional Training Grant award T32DA07242 from NIDA, and Centers of Biomedical Research Excellence P20GM103644 award from the National Institute on General Medical Sciences Abuse. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Food and Drug Administration.

Competing Interests:

The authors have no conflicts of interest to report.

References

- Asfar, T., Arheart, K.L., Dietz, N.A., Caban-Martinez, A.J., Fleming, L.E., Lee, D.J. (2015). Changes in cigarette smoking behavior among US young workers 2005-2010: The role of occupation. *Nicotine Tob Res.* Oct 26. pii: ntv240. [Epub ahead of print]
- Ashley, D.L., Backinger, C.L., van Bommel DM, Neveleff DJ. (2014). Tobacco regulatory science: research to inform regulatory action at the Food and Drug Administration's Center for Tobacco Products. *Nicotine Tob. Res.* 16(8):1045-9. doi: 10.1093/ntr/ntu038. Epub 2014 Mar 17. PMID: 24638850.
- Austin, P.C. (2007). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statist. Med.* 26:2937-57. PMID: 17186501.
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.
- Center for Behavioral Health Statistics and Quality. (2014). 2013 National Survey on Drug Use and Health Public Use File Codebook. Rockville, MD: Substance Abuse and Mental Health Services Administration; http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=35509&ds=1&file_id=1166336&path=SAMHDA. Accessed June 16, 2015.
- Chilcoat, H.D. (2009). An overview of the emergence of disparities in smoking prevalence, cessation, and adverse consequences among women. *Drug Alcohol. Depend.* 104 Suppl 1, S17-S23. doi: 10.1016/j.drugalcdep.2009.06.002. Epub 2009 Jul 24.
- Centers for Disease Control and Prevention. (2011). Current cigarette smoking prevalence among working adults-United States, 2004-2010. *MMWR*, 60(38), 1305-1309.
- Cutler, D.M., Lleras-Muney, A. (2010). Understanding differences in health behaviors by education. *J Health Econ.* Jan;29(1):1-28. doi: 10.1016/j.jhealeco.2009.10.003. Epub 2009 Oct 31.

Davis, D.R., Kurti, A.N., Redner, R., White, T.J., Higgins, S.T. (under review). A review of the literature on contingency management in the treatment of substance use disorders, 2009-2015. *Prev. Med.*

Fiore M., et al. (2008). A clinical practice guideline for treating tobacco use and dependence: 2008 update. A U.S. public health service report. *Am. J. Prev. Med.* 35: 158-76. doi: 10.1016/j.amepre.2008.04.009.

Gaalema, D.E., Cutler, A.Y., Higgins, S.T., Ades, P.A. (2015). Smoking and cardiac rehabilitation participation: associations with referral, attendance, and adherence. *Prev Med.* Nov;80:67-74. doi: 10.1016/j.ypmed.2015.04.009. Epub 2015 Apr 18.

Gfroerer, J., King, B.A., Garrett, B.E., Babb, S., McAfee, T. (2013). Vital signs: current cigarette smoking among adults aged ≥ 18 years with mental illness—United States, 2009-2011. *MMWR*, 62, 81-87.

Graham, H., Inskip, H.M., Francis, B., Marman, J. (2007). Pathways of disadvantage and smoking careers: evidence and policy implications. *Journal of Epidemiology and Community Health (Suppl II)*. 60: ii7-ii12.

Higgins, S.T. (2014). Behavior change, health, and health disparities: An introduction. *Prev. Med.* Nov;80:1-4. doi: 10.1016/j.ypmed.2015.07.020. Epub 2015 Aug 6.

Higgins S.T., Chilcoat H.D. (2009). Women and smoking: an interdisciplinary examination of socioeconomic influences. *Drug Alcohol Depend.* 104 Suppl 1: S1-5. doi: 10.1016/j.drugalcdep.2009.06.006. Epub 2009 Jul 8.

Higgins, S.T., Heil, S.H., Badger, G.J., Skelly, J.M., Solomon, L.J., & Bernstein, I.M. (2009). Educational disadvantage and cigarette smoking during pregnancy. *Drug Alcohol Depend.* October 1; 104(suppl 1): S100-S105. Doi:10.1016/j.drugalcdep.2009.03.013

Higgins, S.T., Kurti, A.N., Redner, R., White, T.J., Gaalema, D.E., Roberts, M.E., Doogan, N.J., Tidey, J.W., Miller, M., Stanton, C.A., Henningfield, J.E., & Atwood, G.S. (2015). A review of the

literature on prevalence of gender differences and intersections with other vulnerabilities to tobacco use in the United States, 2004-2014. *Prev. Med.* Jun 26. pii: S0091-7435(15)00206-6. doi:

10.1016/j.ypmed.2015.06.009. [Epub ahead of print] PMID: 26123717

Hiscock, R., Bauld, L., Amos, A., Fidler, J.A., Munafo, M. (2012). Socioeconomic status and smoking: a review. *Ann NY Acad Sci* Feb;1248:107-23. doi: 10.1111/j.1749-6632.2011.06202.x. Epub 2011 Nov 17.

Hser, Y.I., Hoffman, V., Grella, C.E., & Anglin, M.D. (2001). A 33-year follow-up of narcotics addicts. *Arch Gen Psychiatry*, May;58(5):503-8.

Kandel, D.B., Griesler, P.C., Schaffran, C. (2009). Educational attainment and smoking among women: risk factors and consequence for offspring. *Drug Alcohol Depend.* Oct 1;104 Suppl 1:S24-33. doi: 10.1016/j.drugalcdep.2008.12.005. Epub 2009 Jan 28.

Lei, Y., Nollen, N., Ahluwalia, J.S., Yu, Q., & Mayo, M.S. (2015). An application in identifying high-risk populations in alternative tobacco product use utilizing logistic regression and CART: A heuristic comparison. *BMC Public Health*, 15: 341. doi: 10.1186/s12889-015-1582-z.

Lemon, S.C., Roy, J., Clark, M.A., Friedmann, P.D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Ann Behav Med*, 26(3): 172-181.

Niaura, R., Chander, G., Hutton, H., & Stanton, C. (2012). Interventions to address chronic disease and HIV: strategies to promote smoking cessation among HIV-infected individuals. *Current HIV/AIDS Rep.* Dec; 9(4): 375-84. doi: 10.1007/s11904-012-0138-4.

Office of Disease Prevention and Health Promotion (2015). Healthy People 2020. U.S. Department of Health and Human Services, Washington, DC. <http://www.healthypeople.gov>, accessed November 23rd, 2015.

Park Y., Freedman A.N., Gail, M.H., Pee, D., Hollenback, A., Schatzkin A., Pfeiffer R.M. (2009). A colorectal cancer risk prediction tool for white men and women without known susceptibility. *J Clin Oncol* 27(5):694–698. doi: 10.1200/JCO.2008.17.4813. Epub 2008 Dec 29.

Population Assessment of Tobacco and Health (PATH). (2016).
<https://pathstudyinfo.nih.gov/UI/HomeMobile.aspx>. Accessed February 13, 2016.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>. Accessed 11/7/2015.

Redner, R., White, T.J., Harder, V.S., Higgins, S.T. (2014a). Vulnerability to smokeless tobacco use among those dependent on alcohol or illicit drugs. *Nicotine Tob Res.* 16:216-23. doi: 10.1093/ntr/ntt150. Epub 2013 Sep 30. PMID: 24081975.

Redner, R., White, T.J., Harder, V.S., Higgins, S.T. (2014b). Examining vulnerability to smokeless tobacco use among adolescents and adults meeting diagnostic criteria for major depressive disorder. *Exp Clin Psychopharmacol.* 22:316-22. doi: 10.1037/a0037291. Epub 2014 Jun 30. PMID: 24978349

Rowa-Dewar, N., Lumsdaine, C., Amos, A. (2015). Protecting children from smoke exposure in disadvantaged homes. *Nicotine Tob Res.* 17 (4): 496-501. doi: 10.1093/ntr/ntu217.

Schnohr, P., Scharling, H., Nordestgaard, B.G. (2002). Coronary heart disease risk factors ranked by importance for the individual and community. *European Heart Journal*, 23(8): 620-626. PMID: 11969276.

Schroeder, S.A. (2007). Shattuck Lecture: We can do better—improving the health of the American people. *N Engl J Med.*, Sep 20; 357(12):1221-8.

Schroeder, S.A., Koh, H.K. (2014). Tobacco control 50 years after the 1964 surgeon general's report. *JAMA.* Jan 8;311(2):141-3. doi: 10.1001/jama.2013.285243.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2012). Results from the 2011 National Survey on Drug Use and Health: Summary of the National Findings, NSDUH Series H-44 (HHS Publication No. (SMA) 12-4713). Rockville, MD.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2013). Results from the 2012 National Survey on Drug Use and Health: Summary of the National Findings, NSDUH Series H-46 (HHS Publication No. (SMA) 13-4795). Rockville, MD.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2014). Results from the 2013 National Survey on Drug Use and Health: Summary of the National Findings, NSDUH Series H-46 (HHS Publication No. (SMA) 14-4863). Rockville, MD.

Therneau, T., Atkinson, B., & Ripley, B. (2013). rpart: Recursive Partitioning. R package version 4.1-3. <http://CRAN.R-project.org/package=rpart>.

Tidey, J.W. (under review). A behavioral economic perspective on smoking persistence in serious mental illness. *Prev. Med.*

U.S. Department of Health and Human Services, 2012. The Health Consequences of Smoking: Preventing Tobacco Use Among Youth and Young Adults. A Report of the Surgeon General. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.

White, T.J., Redner, R., Bunn, J.Y., & Higgins, S.T. (2015). Do socioeconomic risk factors for cigarette smoking extend to smokeless tobacco use? *Nic Tob Res.* Oct 26. pii: ntv199. Epub ahead of print. PMID: 26503735.

TABLE 1. Prevalence of current smoking and results from univariate and multivariate logistic regressions predicting current cigarette smoking^a among adults (aged ≥ 18 years) across eight potential risk factors ($n = 114,426$) — National Survey on Drug Use and Health (NSDUH), United States, 2011–2013.

	% Current Smoking		Univariate Logistic Regression		Multivariate Logistic Regression	
	Prevalence		Main effects		Main effects	
	%	(95% CI)	OR	(95% CI)	AOR	(95% CI)
Overall	21.6	(21.1, 22.1)				
Gender						
Male	24.3	(23.6, 25.0)	1.4**	(1.3, 1.4)	1.3**	(1.3, 1.4)
Female	19.0	(18.5, 19.5)	Ref. group		Ref. group	
Age Group (years)						
18 – 25	24.7	(24.2, 25.1)	3.1**	(2.8, 3.5)	2.5**	(2.2, 2.8)
26 – 44	27.0	(26.2, 27.7)	3.5**	(3.1, 3.9)	4.1**	(3.6, 4.6)
45 – 64	21.3	(20.4, 22.1)	2.6**	(2.3, 2.9)	2.8**	(2.5, 3.2)
≥ 65	9.6	(8.6, 10.5)	Ref. group		Ref. group	
Race/Ethnicity^b						
White	23.4	(22.8, 24.0)	3.2**	(2.8, 3.7)	2.8**	(2.4, 3.3)
Black	22.1	(20.9, 23.3)	3.0**	(2.5, 3.5)	1.8**	(1.5, 2.2)
Hispanic	15.4	(14.6, 16.3)	1.9**	(1.7, 2.2)	0.9	(0.8, 1.1)
American Indian/Alaska Native	37.2	(31.2, 43.1)	6.2**	(4.6, 8.4)	3.0**	(2.2, 4.2)
Asian	8.7	(7.5, 9.8)	Ref. group		Ref. group	
Other	31.0	(27.9, 34.0)	4.7**	(3.9, 5.7)	3.3**	(2.7, 4.1)
Education Level						
< High school	30.8	(29.6, 32.0)	3.7**	(3.4, 4.0)	4.6**	(4.2, 5.1)
High school graduate	26.7	(25.9, 27.5)	3.0**	(2.8, 3.2)	3.4**	(3.1, 3.6)
Some college	23.0	(22.1, 23.8)	2.5**	(2.3, 2.7)	2.5**	(2.3, 2.7)
College graduate	10.8	(10.1, 11.5)	Ref. group		Ref. group	
Poverty Status^c						
Below poverty level	32.8	(31.4, 34.1)	2.0**	(1.9, 2.1)	1.6**	(1.5, 1.7)
At or above poverty level	19.6	(19.1, 20.2)	Ref. group		Ref. group	
Any Mental Illness^d						
Yes	31.7	(30.7, 32.7)	1.9**	(1.9, 2.0)	1.5**	(1.4, 1.6)
No	19.2	(18.7, 19.7)	Ref. group		Ref. group	
Alcohol Abuse/Dependence^e						
Yes	44.3	(42.8, 45.9)	3.2**	(3.0, 3.4)	2.3**	(2.1, 2.5)
No	19.8	(19.3, 20.3)	Ref. group		Ref. group	
Illicit Drug Abuse/Dependence^e						
Yes	63.7	(61.4, 66.0)	6.8**	(6.1, 7.6)	3.7**	(3.2, 4.2)
No	20.5	(20.0, 21.0)	Ref. group		Ref. group	

** $p < 0.001$.

Notes. OR = Odds ratio, AOR = Adjusted odds ratio, CI = Confidence interval, Ref. group = Reference group.

^aPersons who reported ever smoking all or part of a cigarette in the 30 days preceding the interview AND smoked ≥ 100 cigarettes in their lifetime. ^bThe five racial/ethnicity categories (White, Black, Hispanic, American Indian/Alaska Native, Asian, Other) are mutually exclusive; “Other” includes Native Hawaiians or Other Pacific Islanders and persons of two or more races. Persons identified as Hispanic might be of any race. ^cBased on reported family income and poverty thresholds published by the U.S. Census Bureau. ^dAny mental illness is defined by the NSDUH as a diagnosable mental, behavioral, or emotional disorder, other than a developmental or substance use disorder, that met the criteria found in the 4th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV). For details on the methodology, see Section B.4.3 in Appendix B of the Results from the 2011 NSDUH: Mental Health Findings. ^eDrug and alcohol abuse and dependence criteria used in the NSDUH were defined based upon the criteria listed in the DSM-IV. Illicit substances included marijuana, hallucinogens, heroin, inhalants, tranquilizers, cocaine, pain relievers, stimulants, and sedatives.

TABLE 2. Results from multiple logistic regression predicting current cigarette smoking^a examining significant main effects and interactions across eight risk factors considered in combination ($n = 114,426$). — National Survey on Drug Use and Health (NSDUH), United States, 2011–2013.

Main Effects	Wald χ^2	p value
Gender	38.22	< 0.0001
Age Group (years) ^b	3.62	0.306
Race/Ethnicity ^c	17.42	0.004
Education Level ^d	17.66	0.001
Poverty Status ^e	1.59	0.208
Any Mental Illness ^f	15.15	< 0.0001
Alcohol Abuse/Dependence ^g	543.57	< 0.0001
Drug Abuse/Dependence ^g	353.84	< 0.0001
Interactions	Wald χ^2	p value
Race/Ethnicity ^c * Age Group (years) ^b	138.24	< 0.0001
Race/Ethnicity ^c * Education Level ^d	292.01	< 0.0001
Race/Ethnicity ^c * Poverty Status ^e	33.58	< 0.0001
Race/Ethnicity ^c * Any Mental Illness ^f	31.71	< 0.0001
Race/Ethnicity ^c * Gender	125.45	< 0.0001
Age Group (years) ^b * Education Level ^d	113.52	< 0.0001
Age Group (years) ^b * Poverty Status ^e	41.92	< 0.0001
Drug Abuse/Dependence ^g * Alcohol Abuse/Dependence ^g	30.85	< 0.0001

^aPersons who reported smoking of all or part of a cigarette in the 30 days preceding the interview AND smoking ≥ 100 cigarettes in their lifetime. ^bAmong persons aged ≥ 18 years (18-25, 26-44, 45-64, ≥ 65 years). ^cThe five racial/ethnicity categories (White, Black, Hispanic, American Indian/Alaska Native, Asian, Other) are mutually exclusive; "Other" includes Native Hawaiians or Other Pacific Islanders and persons of two or more races. Persons identified as Hispanic might be of any race. ^d< HS, HS, some college, college graduate. ^eBased on reported family income and poverty thresholds published by the U.S. Census Bureau. ^fAny mental illness is defined by the NSDUH as a diagnosable mental, behavioral, or emotional disorder, other than a developmental or substance use disorder, that met the criteria found in the 4th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV). For details on the methodology, see Section B.4.3 in Appendix B of the Results from the 2011 NSDUH: Mental Health Findings. ^gDrug and alcohol abuse and dependence criteria used in the NSDUH were defined based upon the criteria listed in the DSM-IV. Illicit substances included marijuana, hallucinogens, heroin, inhalants, tranquilizers, cocaine, pain relievers, stimulants, and sedatives.

Figure Legends

Figure 1. Outcomes of two-way interaction testing among significant risk factors for current smoker status in the multiple logistic regression analysis; + and - symbols indicate risk-factor combinations where there was and was not a significant interaction, respectively.

Figure 2. Three illustrative examples of significant two-way interactions of risk factors for current smoker status; data points represent odds ratios.

Figure 3. A pruned, weighted classification and regression tree (CART) model of associations between current (past 30 days) smoking status and the following eight risk factors in the U.S. adult (≥ 18 years of age) population: educational attainment, age, race/ethnicity, past year drug abuse/dependence, past year alcohol abuse/dependence, annual income below federal poverty level, and past year mental illness. Results from a saturated model were “pruned” using CART analytic software to reduce complexity (R Development Core Team, 2008). Rectangles (nodes) represent smoking prevalence rates for the entire population (top-most node) or population subgroups (all others nodes). Nodes also list the proportion of the adult population represented. Using the root node as an example, 78 percent of the population are non-smokers, 22% smokers, and this node represents 100 of the U.S. non-institutionalized adult population. Lines below nodes represent the binary “yes”-“no” branching around particular risk factors and risk-factor levels, with subgroups in whom the risk factor/level is present moving leftward and downward and those in whom it is absent moving rightward and downward for further potential partitioning based on additional risk factors/levels. The bottom row comprises terminal nodes (i.e., final partitioning for a particular subgroup). Note that minimal terminal node size was set to $\geq 1,000$

individuals. Terminal nodes contain the same information as the other nodes plus the percent of all adult current smokers represented by that node. Percent of current smokers represented is calculated by the following equation: % total population represented by a node X smoking prevalence in that node / smoking prevalence in the entire study sample X 100. Tallying % current smokers represented across all terminal nodes should = 100% of smokers in the U.S adult population save possible rounding error.

Figure 1.

Current Smokers

	Mental Illness	Poverty	Drug	Gender	Education	Age	Alcohol	Race
Race	X	X	-	X	X	X	-	
Alcohol	-	-	X	-	-	-		
Age	-	X	-	-	X			
Education	-	-	-	-				
Gender	-	-	-					
Drug	-	-						
Poverty	-							
Mental Illness								

Figure 2

Figure 2

Current Smoking - Selected Interactions

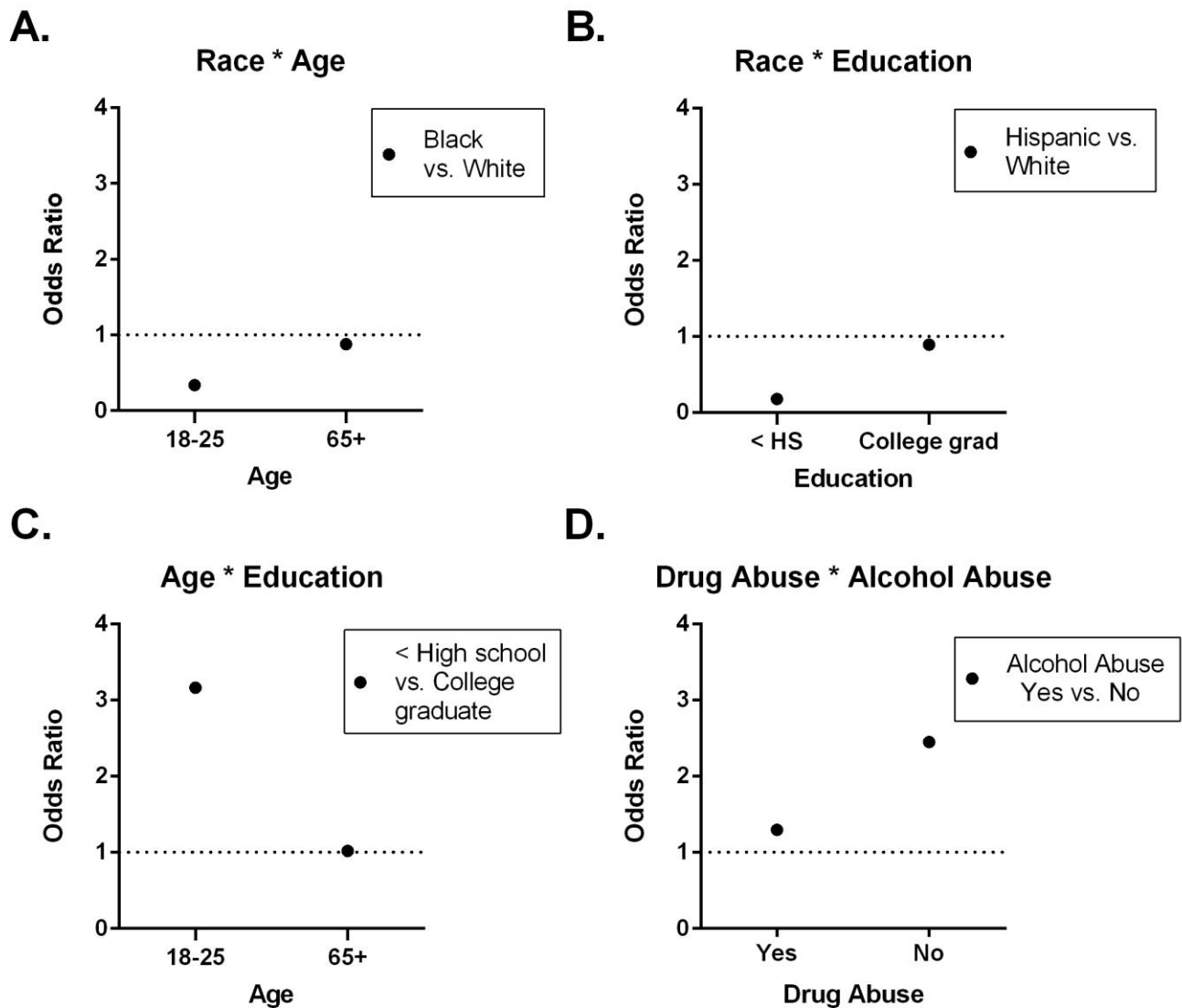
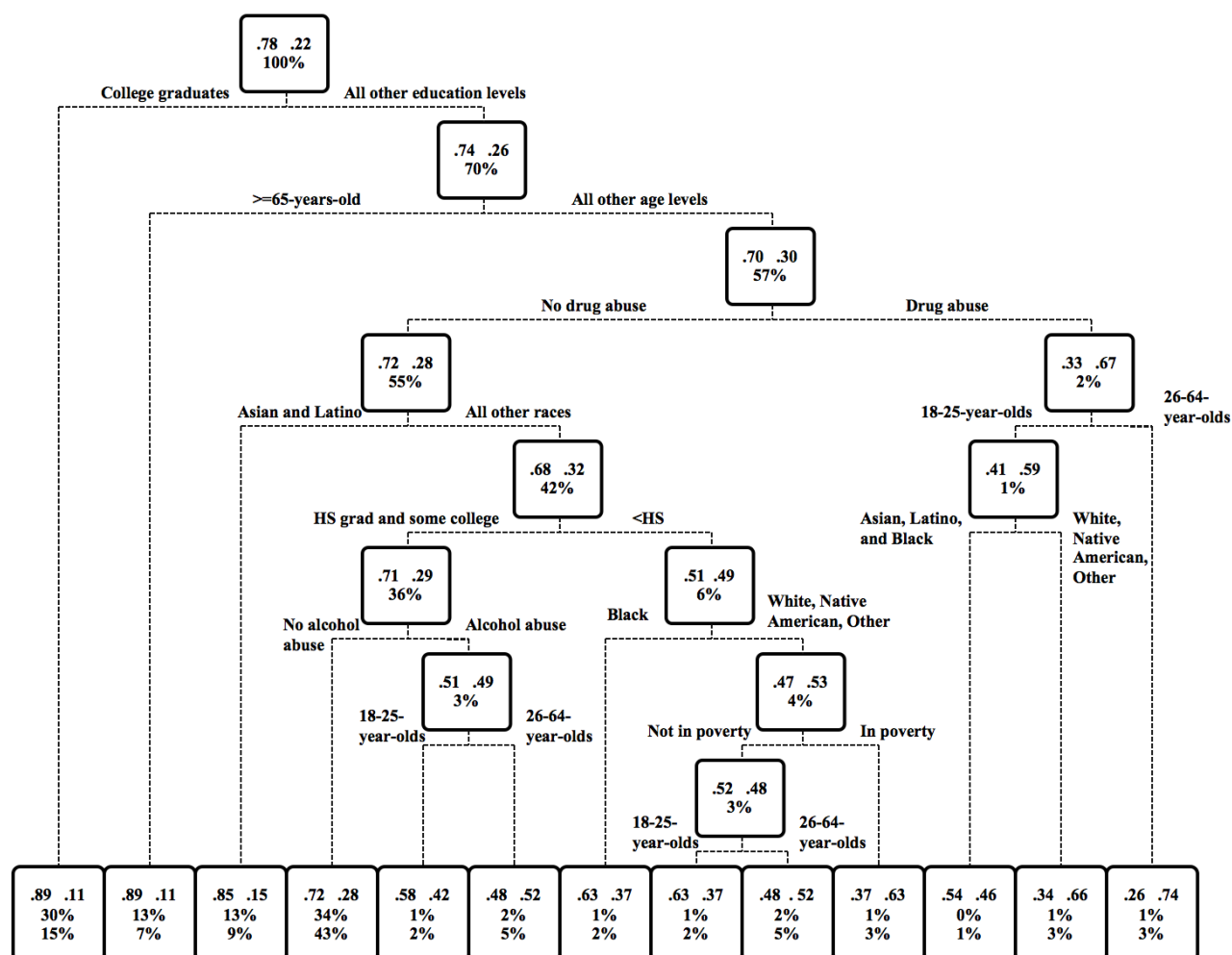


Figure 3



Highlights

Study of co-occurring risk factors for cigarette smoking in U.S. adults

Eight common risk factors each independently predicted current smoking

Educational attainment was the single strongest predictor of smoking status

Effects of risk factors for smoking are often independent, cumulative and summative

ACCEPTED MANUSCRIPT