Using Data Fusion with Multiple Imputation to Correct for Misclassification in Self-Reported Exposure: A Case-Control Study of Cannabis Use and Homicide Victimization

Abstract

Background: Cannabis use has been causally linked to violent behaviors in experimental and case studies, but its association with homicide has not been rigorously assessed through epidemiologic research.

Methods: We performed a case-control analysis using two national data systems. Cases were homicide victims from the National Violent Death Reporting System (NVDRS), and controls were participants from the National Survey on Drug Use and Health (NSDUH). While the NVDRS contained toxicological testing data on cannabis use, the NSDUH only collected self-reported data, and thus the potential misclassification in the self-reported data needed to be corrected. We took a data fusion approach by concatenating the NSDUH with a third data system, the National Roadside Survey of Alcohol and Drug Use by Drivers (NRS), which collected toxicological testing and self-reported data on cannabis use for drivers. The data fusion approach provided multiple imputations (MIs) of toxicological testing results on cannabis use for the participants in the NSDUH, which were then used in the case-control analysis. Bootstrap was used to obtain valid statistical inference.

Results: The analyses revealed that cannabis use was associated with 3.55-fold (95% CI: 2.75 - 4.35) increased odds of homicide victimization. Alcohol use, Black race, male sex, 21-34 years of age, and having less than high school in education attainment were also significantly associated with increased odds of homicide victimization.

Conclusions: Cannabis use is a major risk factor for homicide victimization. The data fusion

with MI method is useful in integrative data analysis for harmonizing measures between different data sources.

Keywords: alcohol use, cannabis use, integrative data analysis, missing at random, multiple imputation, national surveys, stratified bootstrapping.

INTRODUCTION

Homicide has long been a major public health issue in the United States. It is a leading cause of death for those aged between 5 and 44 during 1980-2019.[1] In 2020, homicide claimed 24,576 lives, yielding a death rate of 7.5 per 100,000 population.[2,3] It is well known that excess alcohol consumption is positively associated with the risk of violence, including homicide, suicide, and sexual assault.[4,5] Almost one in four, or 2.7 million out of the 11.1 million victims of violent crime, report that the offender had been drinking alcohol prior to committing the crime each year.[6] Of the 12,638 homicide victims with toxicological testing results in 9 states between 2004 and 2016, 37.5% tested positive for alcohol, 31.0% positive for cannabis, and 11.4% positive for both substances, and the prevalence of cannabis use detected in homicide victims increased from 22.3% in 2004 to 42.1% in 2016.[7]

Although cannabis use has been causally linked to violent behaviors in experimental studies[8-12] and case studies[13], the association between cannabis use and homicide victimization has not been rigorously assessed through epidemiologic research partly because toxicological testing data for the general population are lacking. In this paper, a case-control analysis aimed at assessing the association between cannabis use and the risk of homicide victimization was conducted using three national data systems, including the 2013-2014 National Violent Death Reporting System (NVDRS), the 2013-2014 National Survey on Drug Use and Health (NSDUH), and the 2013-2014 National Roadside Survey of Alcohol and Drug Use by Drivers (NRS). Cases were homicide victims recorded in the NVDRS and controls were participants in the NSDUH, a nationally representative sample. The exposure of primary interest was cannabis use, which was measured based on toxicological testing of blood specimens for the cases and self-report for the controls. Because self-reported cannabis use data might be more susceptible

than toxicological testing to misclassification error, a case-control analysis directly comparing the NVDRS and NSDUH is neither feasible nor valid. In order to address this problem, we devised a data fusion with multiple imputation approach to correct for the misclassification error in the self-reported cannabis use data in the NSDUH by borrowing the data from the NRS, which contained both toxicological testing and self-reported data on cannabis use for a purposeful sample of drivers. To obtain valid statistical inference in the case-control analysis of the imputed NVDRS and NSDUH, we used stratified bootstrap inference with multiple imputation.[14-16]

METHODS

Study population and data collection

National Violent Death Reporting System. The NVDRS is a population-based surveillance system that collects data from participating states in the US regarding violent deaths obtained from death certificates, coroner/medical examiner reports, law enforcement reports, and toxicology reports.[17] The 2013 NVDRS included data from 17 states (Alaska, Colorado, Georgia, Kentucky, Maryland, Massachusetts, North Carolina, New Jersey, New Mexico, Ohio, Oklahoma, Oregon, Rhode Island, South Carolina, Utah, Virginia, and Wisconsin), a total of 18,765 fatal incidents involving 19,251 deaths.[18] The reporting system included information about decedent demographic characteristics, whether alcohol and substance tests were positive, manner of death, and month in which the death occurred. In this study, the 2013 of 4,110 homicide victims aged 16 or older were included.

National Survey on Drug Use and Health. The NSDUH provides nationally representative survey data that contains information about the use of illicit drugs, alcohol, and tobacco among members of the US civilian, noninstitutionalized population aged 12 or older in

all 50 states and the District of Columbia. Survey samples were selected using a stratified multistage sampling design and data were weighted to be representative of the US general population.[19] The survey included questions on respondent demographic characteristics, alcohol and substance use, and mental health. In this study, the 2013 data of 43,365 survey participants aged 16 or older were included.

National Roadside Survey of Alcohol and Drug Use by Drivers. The NRS is designed to gauge alcohol and drug use by drivers on the US roadways and has been conducted in 1973, 1986, 1996, 2007, and 2013, with non-alcohol drug data being included in the 2007 and 2013 surveys. Participants in the 2013 NRS were non-commercial drivers randomly selected at 300 locations across the 48 contiguous states during designated time segments (9:30 am to 11:30 am and 1:30 pm to 3:30 pm on Fridays and from 10 pm to midnight and 1 am to 3 am on both Friday and Saturday nights).[20] The sample was selected using a multistage sampling method and survey weights were provided to make the sample representative of the U.S. driver population.[21] The survey included questions on driver demographic characteristics, drinking and substance use, and trip and vehicle information. In addition, breath alcohol, oral fluid alcohol, and oral fluid drug concentration tests were administered and whole blood specimens were collected during the survey process. In this study, the 2013 data of 11,314 drivers aged 16 and older were included.

Cannabis and alcohol use

In the NVDRS, alcohol and cannabis positivity were determined based on blood sample tests. Blood alcohol concentrations (BACs) were measured in grams per deciliter, and BACs of 0.01 g per deciliter or greater were considered alcohol positive.[22] Cannabis blood tests provided binary results - whether the test for cannabis was positive or negative. Differently, self-reported alcohol

and cannabis use were recorded as binary variables in the NSDUH, indicating whether one used alcohol or cannabis in the past month. Finally, alcohol and cannabis use were measured in both blood and oral fluid samples and asked with questionnaires in the NRS. Both the BACs and the raw alcohol levels in oral fluids were measured in milligrams per deciliter, and a value of 10 mg (or 0.01 g) per deciliter or greater was considered alcohol positive. Blood and oral cannabis tests indicated whether the test for Tetrahydrocannabinol was positive or negative. The questions about the last time of cannabis use (past 24 hours, past 2 days, past month, over a month, beyond a year/never) and the frequency of weekly alcohol consumption (0, 1-2, 3-4, 5-7, 8-14, more than 14, did not answer) were asked. We created the binary self-reported alcohol and cannabis use variables to be comparable with the self-reported variables in the NSDUH. Specifically, individuals who reported cannabis use in the past 24 hours, past 2 days, or past month were categorized as cannabis use in the past month, and no use otherwise; and individuals who reported with non-zero alcohol consumption frequency were categorized as alcohol use in the past month, and no use otherwise.

Covariates

The variables included as covariates in the case-control analysis are presented in Table 1. Demographic variables include age (16-20 years, 21-34 years, 35-49 years, 50-64 years, or ≥65 years), sex (male vs. female), race (White, Black, Hispanic, or Others), and education (less than high school, high school graduate, some college, or college graduate/some graduate). All the covariates are fully measured in the NSDUH but have missing values in both the NRS and the NVDRS.

Statistical analysis

We performed the descriptive analysis to compare the distributions of the cannabis and alcohol use variables as well as the covariates between the three data systems. For the NRS and NSDUH, frequencies and the survey weighted percentage were provided; while for the NVDRS, frequencies and the unweighted percentage were reported. To assess the degree of misclassification associated with the self-reported results, we used the NRS data with complete pairs of self-reported use and blood test for both cannabis and alcohol to calculate the sensitivity and specificity of self-reported cannabis and alcohol use.

To correct for potential misclassification in the self-reported cannabis and alcohol use in the NSDUH, we devised a data fusion approach. We first concatenated the NSDUH with the NRS. The oral and blood test results of cannabis and alcohol use in the NSDUH were treated as missing data. We then performed multiple imputation (MI) to fill in the missing test results in the NSDUH using the Chained Equations Multiple Imputation (CEMI) algorithm[23-24] assuming data are missing at random. The CEMI algorithm is an iterative procedure, imputing one variable at a time conditioning on all the other variables in the data set. To impute the missing binary test result of cannabis and alcohol use, we considered three different imputation models, including logistic regression, lasso logistic regression, and random forest. We compared the performance of the three imputation models using area under curve, sensitivity, and specificity with 10-fold cross-validation and found that the lasso logistic model yielded the highest area under curve (Supplementary eFigure 1). Hence, the lasso logistic regression was chosen as the imputation models in the imputation of test results of cannabis and alcohol use. The CEMI algorithm was implemented using the "mice" package[25] in R and created multiple imputations of blood test results of cannabis and alcohol use in the NSDUH, which were then used in the next step of the

analyses.

Some covariates were missing in the NVDRS. We multiply imputed the missing data in the NVDRS again using the "mice" package in R by assuming data are missing at random and created the same number of imputations as in the NSDUH. We then concatenated each imputation of the NSDUH with each imputation of the NVDRS and created combined imputed NSDUH and NVDRS data sets. For the purpose of modeling with the NSDUH data in the integrative data analysis, survey design variables were manually added to the NVDRS. Specifically, a unique value was assigned to each observation for the primary sampling unit variable; all observations were assigned with a single stratum value which is different from the strata values in the NSDUH; and the value 1 was assigned to each observation for the weight variable.

We fit weighted logistic regression models on the multiply imputed integrative data of the NSDUH and NVDRS, accounting for the stratified multistage sampling design in the NSDUH. Because the NRS was involved in the imputation but was not used in the post-imputation case-control analysis, the conventional variance estimation method for multiply imputed data using the Rubin's MI combining rules[26] does not apply anymore.[14] Yu et al. showed that the bootstrapping with multiple imputation (BMI) yields valid statistical inference in this setting.[15] The BMI method is implemented by first bootstrapping the samples and then conducting MI within each bootstrapped sample.[16] To obtain bootstrap samples, resampling was conducted on primary sampling units within each stratum for the NSDUH and NRS, and on homicide victims for the NVDRS. We conducted 200 bootstraps with 2 imputations within each bootstrapped combined sample. This yields 400 multiply imputed integrative data sets of the NSDUH and NVDRS. We fit one weighted logistic regression model on each imputed

integrative data set. The results of 400 sets of regression coefficients were then pooled to obtain the final point estimates of the regression coefficients and bootstrap variance estimates that account for the variation between the multiple imputations within a bootstrapped data set, and between the bootstrapped data sets. The 95% confidence intervals (CI) of the regression coefficients are then obtained using a t-distribution with degrees of freedom computed by Satterthwaite approximation.[16] We called it model 1.

For comparison, we also fit the other three weighted logistic multivariable models. Model 2 was fitted using 20 multiply imputed NVDRS and NSDUH data after the process of data fusion with MI, with the variance of the regression coefficients estimated using the Rubin's method for MI combination. Model 2 was included to show the difference in the 95% CI interval estimates using the Rubin's method versus the BMI method. Models 3 and 4 considered a different case-control analysis using the participants in the NRS as the controls. Different from the NSDUH in which the participants represented the general U.S. population, the participants in the NRS only represented the drivers in the US. Therefore, the case-control analysis using the NRS as the controls is less desirable than that using the NSDUH as the controls. Because the blood test results of cannabis and alcohol use were available in the NRS, data fusion with MI was not required in the case-control analysis using the NRS as controls. Model 3 was fitted using the complete cases of the NVDRS and NRS data using listwise deletion of any incomplete observations. Model 4 was fitted using 20 multiply imputed NVDRS and NRS data, with the incomplete covariates imputed using the "mice" package in R by assuming data are missing at random and the variance of the regression coefficients estimated using the Rubin's method for variance estimation.

RESULTS

Table 1 shows the distributions of the demographic variables as well as the cannabis and alcohol use variables in the NVDRS, NSDUH, and NRS. The 2013 NSDUH is a representative sample of the US general population, and the NRS is a representative sample of drivers only. Among individuals aged 16 years and older, compared to the US general population, drivers had higher percentage of males (58.3% vs. 48.2%), higher percentage of people aged between 21 and 34 (39.8% vs. 24.0%), lower percentage of White population (55.0% vs. 65.5%) but higher percentage of Black population (24.4% vs. 11.8%), higher percentage of some college or college education (69.1% vs. 58.4%), and higher percentage of self-reported cannabis use (11.7% vs. 7.9). On the other hand, there was much higher percentage of males (81.3%), Black people (56.1%), people with education less than or equal to high school (83.2%) among the homicide victims. It is also noticeable that 46.8% of the homicide victims had a positive blood cannabis test and 44.7% had a positive blood alcohol test; while only 11.0% and 2.7% of drivers had positive blood cannabis and alcohol test, respectively.

Table 2 shows the results of the sensitivity/specificity analyses for the self-reported cannabis and alcohol use with the blood test results as the gold standard using the complete pairs of self-reported and blood test results in the NRS. The self-reported cannabis use had a high specificity (94%) but low sensitivity (62%); whereas the self-reported alcohol use had a high sensitivity (88%) but low specificity (47%). This finding is expected because cannabis stays longer in the body than alcohol and the self-reported variables are based on the past month use. In addition, the sensitivity and specificity varied between different age, gender, race, and education subgroups. For example, the Black race group had lower sensitivity than the White race group in both the self-reported cannabis (56% vs. 66%) and alcohol (71% vs. 95%) use.

These results indicate the existence of misclassification error in the self-reported cannabis and alcohol use variables and hence the correction for the misclassification is necessary. Further the misclassification is differential with the degree of misclassification related to demographic variables. These findings highlight the importance of correcting for misclassification error in the self-reported data in the NSDUH and the necessary in accounting for the covariates in the misclassification error adjustment.

The Supplementary eTable 1 repeated the analyses in Table 1 but showed the average estimates based on 20 imputations for each of the three data systems. The distributions of the demographic characteristics are like those in Table 1 using the complete cases. For NSDUH, the data fusion with MI estimated an oral alcohol positivity rate of 2.1% (compared to 2.7% among drivers), a blood alcohol positivity rate of 2.6% (compared to 3.0% among drivers), an oral cannabis positivity rate of 6.9% (compared to 9.8% among drivers), and a blood cannabis positivity rate of 7.4% (compared to 9.9% among drivers).

Table 3 shows the odds ratio (OR) and 95% confidence interval (CI) of the case-control analyses. According to model 1, cannabis use was associated with a 3.55-fold increase (95% CI: 2.75, 4.35) in the odds of homicide victimization. Alcohol use was also strongly associated with the homicide victimization with OR = 19.25 (95% CI: 12.25, 26.24), followed by being racially Black with OR = 5.27 (95% CI: 4.55, 6.0), being male with OR = 2.94 (95% CI: 2.59, 3.3), being aged 21 to 34 years old with OR = 1.79 (95% CI: 1.33, 2.25), and having education less than high school with OR = 1.68 (95% CI: 1.44, 1.91). Comparing model 1 to model 2, the point estimates of the regression coefficients are similar, but the BMI method used in model 1 yielded narrower confidence intervals for the coefficients of the drug variables than the Rubin's method, because the latter overestimated the variance of the regression coefficients in this setting.

Comparing model 1 to model 4, using drivers as the controls did not lead to much difference in the estimated associations between homicide victimization and the use of cannabis or alcohol, but there are notable differences in the ORs associated with the demographic variables. Finally, comparing model 3 to model 4, the MI led to a much larger analyzable data set, with the sample size increased by 3 times. Model 4 estimated a stronger association between homicide victimization and the cannabis and alcohol use and narrower confidence intervals than model 3. The comparison between models 3 and 4 highlights the limitations of complete-cases analysis, where listwise deletion of any incomplete observations could largely reduce statistical power and result in biased statistical inference. MI can be used to overcome these limitations.

DISCUSSION

This study is among the first to assess the association between cannabis use and homicide victimization using rigorous statistical analytic inference and multiple national data systems. Because the data source for cases (i.e., the NVDRS) contained toxicological testing results for cannabis and alcohol use but the data source for controls (i.e., the NSDUH) collected self-reported data only, it is impossible to perform a valid case-control analysis without the employment of the data fusion with multiple imputation method. The data fusion with multiple imputation method we proposed in this paper has wide applications and could facilitate rigorous research using existing data systems to address challenging and important questions.

In this population-based case-control study, we demonstrate that there is a strong association between cannabis use and homicide victimization. The finding that cannabis use is associated with increased odds of being homicide victim is consistent with previous studies.[27-29] Additionally, the estimated associations of alcohol use and demographic characteristics with

homicide victimization obtained by the data fusion with MI method are in line with other reports.[28] This indicates that data fusion with multiple imputation is useful for integrative data analysis in the context of misclassification and missing values. It is interesting to find that the association between cannabis use and homicide victimization was similar between the analyses using the NSDUH as the controls and using the NRS as the controls, even though one represented the US general population and the other represented the driver populations.

Because the data used for imputations (NRS + NSDUH) in the data fusion step differ from the data used in the post-imputation analyses (NVDRS + NSDUH), Rubin's method for MI variance estimation may not yield valid statistical inference and no longer be applicable. We applied the bootstrapping with multiple imputation method for the variance estimation instead. Our analysis shows that the 95% CIs of the regression coefficients estimated using the bootstrapping with MI method are narrower than those estimated using the Rubin's method applied to MI only. This result agrees with the findings in other studies.[14]

The study has some limitations. The questions about cannabis and alcohol use in the NRS were not framed exactly same as in the NSDUH. Specifically, cannabis use was asked about the last time of use (past 24 hours, past 2 days, past month, over a month, beyond a year/never) in the NRS but a binary variable indicating whether one used cannabis in the past month was asked in the NSDUH. For alcohol use, the frequency of weekly alcohol consumption was asked in the NRS but a binary variable indicating whether one used alcohol in the past month was recorded in the NSDUH. We created binary variables of cannabis and alcohol use from the questions asked in the NRS to make the self-reported cannabis and alcohol use variables comparable between the two data systems, but this is less ideal than if the questions were framed the same. Another limitation of our study is that the computation involved in the bootstrapping with MI is intensive.

In this study, we have tried 200 bootstraps with 2 imputations, 200 bootstraps with 3 imputations, and 300 bootstraps with 3 imputations. There was little difference in the estimates. Therefore, we recommend a combination of 200 bootstraps with 2 imputations per bootstrap, which was also recommended by the literature.[16]

Nevertheless, the data fusion with MI method described in this study appears to be a powerful tool for integrative data analysis in epidemiologic studies. Its capacity to harness and harmonize data on variables across multiple data sources could vastly bolster the utility of individual data systems and open new horizons for epidemiologists. Our application of this novel method to three national data systems reveals a robust association between cannabis use and significantly increased risk of homicide victimization while reaffirming the well-documented role of alcohol use and demographic factors in homicide victimization.

REFERENCES

1. Centers for Disease Control and Prevention. Leading causes of death and number of deaths, by age, United States, 1980 and 2019. https://www.cdc.gov/nchs/data/hus/2020-2021/LCODAge.pdf. Reviewed August 12, 2022. Accessed December 10, 2022.
2. Centers for Disease Control and Prevention. Assault or homicide. https://www.cdc.gov/nchs/fastats/homicide.htm. Reviewed December 30, 2022. Accessed December 10, 2022.
3. Centers for Disease Control and Prevention. New CDC/NCHS data confirm largest one-year increase in U.S. homicide rate in 2020. https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2021/202110.htm. Reviewed October 06. 2021. Accessed December 10, 2022.
4. Centers for Disease Control and Prevention. The health effects of excessive alcohol use. https://www.cdc.gov/chronicdisease/resources/publications/factsheets/alcohol.htm#:~:text=Drinking%20too%20much%20alcohol%20increases,from%20opioids%20and%20other%20substances. Reviewed July 11, 2022. Accessed December 10, 2022.
5. National Council on Alcoholism and Drug Dependence. From bar to bars: Links between alcohol and crime. https://www.ncaddesgpv.org/blog/from-bar-to-bars-links-between-alcohol-and-crime. Updated March 31, 2016. Accessed December 10, 2022.
6. Greenfeld LA. Alcohol and crime: An analysis of national data on the prevalence of alcohol involvement in crime. Report prepared for Assistant Attorney General's National Symposium on Alcohol Abuse and Crime. *U.S. Department of Justice,* 1998; NCJ-168632.
7. Nazarov O, Li G. Trends in alcohol and marijuana detected in homicide victims in 9 US states: 2004-2016. *Injury Epidemiology*. 2020;7(1):2. (doi:10.1186/s40621-019-0229-4).

8. Allen RP, Safer D, Covi L. Effects of psychostimulants on aggression. *Journal of Nervous and Mental Disease*. 1975;160:138-145.
9. Alves CN, Carlini EA. Effects of acute and chronic administration of Cannabis sativa extract on the mouse-killing behavior of rats. *Life Sciences*. 1973;13:75-85.
10. Alves CN, Goyos AC, Carlini EA. Aggressiveness induced by marihuana and other psychotropic drugs in REM sleep deprived rats. *Pharmacology Biochemistry and Behavior*. 1973 1:183-189.
11. Beatty WW, Costello KB, Berry SL. Suppression of play fighting by amphetamine: Effects of catecholamine antagonists, agonists and synthesis inhibitors. *Pharmacology Biochemistry and Behavior*. 1984;20:747-755.
12. Beezley DA, Gantner AB, Bailey DS, Taylor SP. Amphetamines and human physical aggression. *Journal of Research in Personality*. 1987;21:52-60.
13. Goldstein, P. J. The drugs/violence nexus: A tripartite conceptual framework. *Journal of Drug Issues*. 1985;*15*(4), 493–506.
14. Reiter JP. Multiple Imputation When Records Used for Imputation Are Not Used or Disseminated for Analysis. *Biometrika*. 2008;*95*(4), 33–946.
15. Yu Y, Little R, Perzanowski M, Chen Q. Multiple Imputation of More than One Environment Exposure with Non-differential Measurement Error. *Biostatistics*, in press.
16. von Hippel PT, Bartlett J. Maximum Likelihood Multiple Imputation: Faster imputations and consistent standard errors without posterior draws. *Statistical Science, Statist. Sci.* 2021;*36*(3), 400-420.
17. Centers for Disease Control and Prevention. National violent death reporting system. https://www.cdc.gov/violenceprevention/datasources/nvdrs/index.html. Reviewed September 28, 2021. Accessed December 10, 2022.
18. Lyons BH, Fowler KA, Jack SP, Betz CJ, Blair JM. Surveillance for Violent Deaths — National Violent Death Reporting System, 17 States, 2013. MMWR Surveill Summ 2016;65(No. SS-10):1–42.
19. National Survey on Drug Use and Health. *Substance Abuse and Mental Health Services Administration, Results from the 2013 National Survey on Drug Use and Health: Mental Health Findings*, Rockville, MD: Substance Abuse and Mental Health Services Administration, 2014. (NSDUH series H-49). (HHS Publication no. (SMA) 14-4887).

20. National Highway Traffic Safety Administration. *Results of the 2013-2014 National Roadside Survey of Alcohol and Drug Use by Drivers.* Washington, DC: National Highway Traffic Safety Administration, 2015. (Report no. DOT HS 812 362).
21. National Highway Traffic Safety Administration. *2013-2014 National Roadside Study of alcohol and drug use by drivers: Drug results.* Washington, DC: National Highway Traffic Safety Administration, 2017. (Report no. DOT HS 812 411).
22. Center for Disease Control and Prevention (CDC). National Violent Death Reporting System: Web Coding Manual. Atlanta: Centers for Disease Control & Prevention (CDC); Revised January 18, 2022. Accessed December 10, 2022.
23. Raghunathan, T. E., Lepkowski, J.M., Hoewyk, J. V. and Solenberger, P.W. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. 2001;27, 85-95.
24. van Buuren, S. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*. 2007;16, 219–242.
25. van Buuren, S. and Groothuis-Oudshoorn, K. Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*. 2011;45, 1-67.
26. Rubin, D. B. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons Inc.; 1987.
27. Hohl BC, Wiley S, Wiebe DJ, et al. Association of drug and alcohol use with adolescent firearm homicide at individual, family, and neighborhood levels. *JAMA internal medicine*. 2017;*177*(3), 317–324.
28. Darke S, Duflou J, Torok M. Drugs and violent death: comparative toxicology of homicide and non-substance toxicity suicide victims. *Addiction.* 2019;104(6), 1000-1005.
29. Nazarov O, Li G. Linking Cannabis and Homicide: Comparison with Alcohol. In: Patel VB, Preedy VR, eds. *Handbook of Substance Misuse and Addictions*. Springer, Cham; 2022:2-12.

Table 1. Distributions of the Baseline Characteristics of the Three US National Data Systems Aged 16 Years and Older: Drivers from the 2013-14 National Roadside Survey of Alcohol and Drug Use (NRS), US population from the 2013-14 National Survey on Drug Use and Health (NSDUH), Homicide Victims from the 2013 National Violent Death Reporting System (NVDRS).

Data Source	NRS n = 11,314	NSDUH n = 43,465	NVDRS n = 4,110
	Frequency (%)	Frequency (%)	Frequency (%)
Age (years)			
16-20	1,069 (11.7)	3,817 (8.8)	489 (11.9)
21-34	3,640 (39.8)	10,414 (24.0)	1,813 (44.1)
35-49	2,310 (25.2)	10,751 (24.7)	997 (24.3)
50-64	1,621 (17.7)	10,755 (24.7)	581 (14.1)
≥65	515 (5.6)	7,728 (17.8)	230 (5.6)
Missing	2,158	0	0
Sex			
Male	6,382 (58.3)	20,970 (48.2)	3,340 (81.3)
Female	4,566 (41.7)	22,495 (51.8)	770 (18.7)
Missing	365	0	0
Race			
White	4,952 (55.0)	28,458 (65.5)	1,229 (29.9)
Black	2,196 (24.4)	5,131 (11.8)	2,306 (56.1)
Hispanic	1,074 (11.9)	6,622 (15.2)	376 (9.1)
Others	776 (8.6)	3,254 (7.5)	199 (4.8)
Missing	2,316	0	0
Education			
Less than high school	715 (7.8)	5,616 (12.9)	887 (38.0)
High school graduate	2,115 (23.1)	12,441 (28.6)	1,055 (45.2)
Some college	3,222 (35.2)	11,365 (26.1)	248 (10.6)
College/Some graduate	3,106 (33.9)	14,043 (32.3)	145 (6.2)
Missing	2,156	0	1,775
Self-reported alcohol use			
Positive	3,811 (58.4)	23,904 (55.0)	-
Negative	2,720 (41.6)	19,561 (45.0)	-
Missing	4,783	0	4,110
Oral alcohol test			
Positive	234 (2.8)	-	-
Negative	8,149 (97.2)	-	-
Missing	2,931	43,465	4,110
Blood alcohol test			
Positive	137 (2.7)	-	1,027 (44.7)
Negative	4,926 (97.3)	-	1,268 (55.3)
Missing	6,251	43,465	1,815
Self-reported cannabis use			
Positive	905 (11.7)	3,417 (7.9)	-
Negative	6,813 (88.3)	40,048 (92.1)	-
Missing	3,596	0	4,110
Oral cannabis test			
Positive	864 (10.3)	-	-
Negative	7,519 (89.7)	-	-
Missing	2,931	43,465	4,110
Blood cannabis test			
Positive	555 (11.0)	-	663 (46.8)
Negative	4,494 (89.0)	-	755 (53.2)
Missing	6,265	43,465	2,692

Table 2. Sensitivity and Specificity of Self-Reported Cannabis and Alcohol Use from the NRS Data Sample with Complete Pairs of Self-Reported Use and Blood Test for Both Cannabis and Alcohol, Stratified by Age, Sex, Race, and Education.

	Cannabis Use			Alcohol Use		
	n	Sens (%)	Spec (%)	n	Sens (%)	Spec (%)
Overall						
All-inclusive	4,221	62	94	3,022	88	47
Age (years)						
16-20	439	66	88	235	100	58
21-34	1,762	63	91	1,441	86	40
35-49	1,039	62	98	744	88	53
50-64	755	44	98	480	91	52
≥65	226	100	99	122	100	52
Sex						
Male	2,361	66	94	1,714	90	40
Female	1,860	56	94	1,308	83	57
Race						
White	2,436	66	94	1,760	95	47
Black	931	56	94	670	71	46
Hispanic	434	70	96	301	90	53
Others	428	62	95	291	100	46
Education						
Less than high school	254	55	96	145	100	54
High school graduate	970	63	95	646	82	53
Some college	1,600	60	93	1,175	84	46
College/Some graduate	1,397	67	95	1,056	95	45

Table 3. Odds Ratio of Homicide Victimization in the Case-Control Study with Victims in the NVDRS as Cases and General Population in the NSDUH or Drivers in the NRS as Controls Using Four Weighted Logistic Multivariable Regression Models: Model 1 Pools the 400 Estimates of Regression Coefficients Using 400 Bootstrapped-then-Imputed NVDRS and NSDUH Data after Data Fusion; Model 2 Pools the 20 Estimates of Regression Coefficients Using 20 Multiply-Imputed NVDRS and NSDUH Data after Data Fusion; Model 3 Shows the Estimate of Regression Coefficients Using the Complete Cases of the NVDRS and NRS Data Using Listwise Deletion; Model 4 Pools the 20 Estimates of Regression Coefficients Using 20 Multiply-Imputed NVDRS and NRS.

	NVDRS + NSDUH		**NVDRS + NRS**	
	Model 1 n = 47,582	Model 2 n = 47,582	Model 3 n = 5,205	Model 4 n = 15,431
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Age (years)				
35-49	Ref	Ref	Ref	Ref
16-20	1.41 (1.08, 1.84)	1.37 (1.06, 1.78)	1.10 (0.67, 1.79)	0.51 (0.38, 0.69)
21-34	1.76 (1.38, 2.25)	1.84 (1.47, 2.29)	1.13 (0.79, 1.64)	0.81 (0.68, 0.95)
50-64	1.37 (1.08, 1.73)	1.46 (1.10, 1.94)	0.83 (0.56, 1.24)	0.79 (0.66, 0.96)
≥65	0.71 (0.54, 0.94)	0.70 (0.53, 0.93)	1.06 (0.51, 2.19)	1.48 (1.05, 2.09)
Sex				
Female	Ref	Ref	Ref	Ref
Male	2.94 (2.60, 3.31)	2.91 (2.59, 3.27)	2.19 (1.65, 2.91)	1.73 (1.45, 2.07)
Race				
White	Ref	Ref	Ref	Ref
Black	5.25 (4.58, 6.02)	5.19 (4.51, 5.97)	2.56 (1.32, 4.96)	2.61 (1.60, 4.27)
Hispanic	0.73 (0.61, 0.87)	0.72 (0.59, 0.87)	2.59 (1.51, 4.42)	0.88 (0.59, 1.30)
Others	1.59 (1.31, 1.93)	1.49 (1.15, 1.93)	1.84 (1.03, 3.28)	1.19 (0.78, 1.80)
Education				
High school graduate	Ref	Ref	Ref	Ref
Less than high school	1.67 (1.45, 1.92)	1.69 (1.45, 1.97)	2.23 (1.68, 2.97)	2.47 (1.97, 3.09)
Some college	0.32 (0.26, 0.38)	0.31 (0.26, 0.38)	0.19 (0.14, 0.26)	0.20 (0.17, 0.25)
College/Some graduate	0.19 (0.16, 0.23)	0.19 (0.16, 0.24)	0.18 (0.12, 0.27)	0.14 (0.11, 0.19)
Drug				
Alcohol (positive vs. negative)	18.53 (12.64, 27.16)	19.09 (11.51, 31.65)	15.83 (9.89, 25.33)	17.99 (12.87, 25.15)
Cannabis (positive vs. negative)	3.50 (2.80, 4.39)	3.59 (2.77, 4.64)	2.58 (1.67, 3.97)	3.53 (2.60, 4.79)