Accident Analysis and Prevention 120 (2018) 262–269

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Multiple imputation of missing marijuana data in the Fatality Analysis Reporting System using a Bayesian multilevel model

Qixuan Chen[a,*], Sharifa Z. Williams[a], Yutao Liu[a], Stanford T. Chihuri[b,c], Guohua Li[b,c,d]

[a] Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY 10032, USA
[b] Center for Injury Epidemiology and Prevention, Columbia University Medical Center, New York, NY 10032, USA
[c] Department of Anesthesiology, Columbia University College of Physicians and Surgeons, New York, NY 10032, USA
[d] Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY 10032, USA

ARTICLE INFO

Keywords:
Bayesian multilevel models
Drugged driving
Fatality Analysis Reporting System
Marijuana use
Missing data
Multiple imputation

ABSTRACT

Background: The Fatality Analysis Reporting System (FARS) provides important data for studying the role of marijuana in motor vehicle crashes. However, marijuana testing data are available for only 34% of drivers in the FARS, which represents a major barrier in the use of the data.

Methods: We developed a multiple imputation (MI) procedure for estimating marijuana positivity among drivers with missing marijuana test results, using a Bayesian multilevel model that allows a nonlinear association with blood alcohol concentrations (BACs), accounts for correlations among drivers in the same states, and includes both individual-level and state-level covariates. We generated 10 imputations for the missing marijuana-testing data using Markov chain Monte Carlo simulations and estimated positivity rates of marijuana in the nation and each state.

Results: Drivers who were at older age, female, using seatbelt at the time of crash, having valid license, or operating median/heavy trucks were less likely to test positive for marijuana. There was a reverse U-shaped association between BACs and positivity of marijuana, with lower positivity when BACs < 0.01 g/dL or ≥ 0.15 g/dL. The MI data estimated a lower positivity rate of marijuana in the nation and each of the state than the observed data, with a national positivity rate of 11.7% (95% CI: 11.1, 12.4) versus 14.8% using the observed data in 2013.

Conclusions: Our MI procedure appears to be a valid approach to addressing missing marijuana data in the FARS and may help strengthen the capacity of the FARS for monitoring the epidemic of drugged driving and understanding the role of marijuana in fatal motor vehicle crashes in the United States.

1. Introduction

In the United States, the prevalence of driving under the influence of non-alcohol drugs has surpassed the prevalence of driving under the influence of alcohol and has become a serious public safety concern (Berning et al., 2015; Compton and Berning, 2015; Romano and Voas, 2011; Li et al., 2013). Marijuana is the most commonly detected non-alcohol drug in drivers. In the past two decades, traffic fatalities involving marijuana have increased markedly, likely resulting from increased permissibility, accessibility and consumption of marijuana. (Berning et al., 2015; Brady and Li, 2014) Although traffic laws may vary from state to state, driving under the influence of drugs (DUID), such as marijuana, is illegal in all states. At least seven US states have per se laws specifying a legal delta-9-tetrahydrocannabinol (THC) threshold for drivers.

The Fatality Analysis Reporting System (FARS), maintained by the National Highway Traffic Safety Administration (NHTSA), contains data derived from an annual census of fatal motor vehicle crashes in the United States and has been frequently used by researchers in assessing crash risk and culpability associated with alcohol and drugs. Despite continued improvement in drug testing, data on drugs are available for only 34% of all drivers involved in fatal crashes and the drug testing rates vary by state and driver characteristics (Brady and Li, 2014; Slater et al., 2016). The missing data represent a major barrier to studying the epidemiologic patterns of marijuana involvement in fatal vehicle crashes, especially in states and subpopulations with low drug testing rates.

The method of multiple imputation (MI) for handling missing data has become increasingly popular in epidemiologic studies because of its improved performance over alternative methods and ease of

* Corresponding author at: Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032, USA.
E-mail address: qc2138@cumc.columbia.edu (Q. Chen).

https://doi.org/10.1016/j.aap.2018.08.021
Received 8 January 2018; Received in revised form 5 July 2018; Accepted 21 August 2018

implementation. (Rubin, 1987; Little and Rubin, 2002) MI proceeds by replacing each missing value with a plausible value and repeating the process multiple times to generate a collection of imputed datasets. The statistical analysis is then performed within each imputed dataset, with final results obtained by combining estimates over the collection of imputed datasets using Rubin's rules, which accounts for both within-imputation and between-imputation variances (Rubin, 1987; Little and Rubin, 2002). In this paper, we proposed a Bayesian multiple imputation procedure to handle missing marijuana data in the FARS. Bayesian models are advantageous in multiple imputation as Bayesian framework automatically and coherently propagate uncertainty about the imputed missing values through the Markov chain Monte Carlo simulation. We then used the multiply-imputed FARS data to estimate positivity rates of marijuana in fatal vehicle crashes in the nation, each census region and state, as well as subgroups of drivers under different crash circumstances. This paper represents one of the first attempts to develop multiple imputations for providing valid and reliable estimates of drug involvement in fatal motor vehicle crashes in the nation and various subpopulations using the FARS data.

2. Methods

2.1. Data source

Data were drawn from the FARS 2013, which is a census of all crashes that occurred on public roads in the United States and resulted in at least one death of an occupant or non-occupant within 30 days of the crash (Analytical, 2012). This dataset includes detailed information on driver characteristics and crash circumstances. We limited our analyses to drivers aged 15 years or older. Our study included 44,518 drivers.

2.2. Toxicological testing data for marijuana

In FARS, drug involvement was recorded by category (e.g., cannabis, narcotics, depressants, and stimulants), and up to three drugs for each driver could be recorded. When four or more drugs were present, narcotics, depressants, stimulants, and cannabis were recorded in a hierarchical order. The cannabis category includes THC, hashish oil, hashish, marijuana, Marinol, and other cannabinoids. We coded marijuana as positive for drivers who tested positive for any substance in the cannabis category; as missing for drivers with missing toxicological testing results or with positive results in narcotics, depressants and stimulants that have higher hierarchy than cannabis (cannabis could get involved but would not be recorded); and as negative otherwise. Among the 44,518 drivers, 15,277 (34.3%) had observed marijuana testing results, 524 (1.2%) were tested but used more than three drugs in which marijuana had lower hierarchy and thus was not recorded, and 28,717 (64.5%) were not tested or whose test results were not registered.

2.3. Covariates

Auxiliary variables included as covariates are described in Table 1, including driver and vehicle characteristics, crash circumstances, and state-level information. The numbers of bus and motor homes are small, so we combined these two groups with miscellaneous vehicles. The distribution of blood alcohol concentrations (BACs) may be regarded as semi-continuous with a substantial portion of BAC values being zero and the remaining continuously spreading across the positive numbers, and thus we created two variables for BACs, with one binary (positive, negative) and the other continuous for blood alcohol concentrations (g/ml). All US states have the same legal alcohol limit for drivers, 0.08 g/dL. Similarly, although enforcement of DUID laws may differ across jurisdictions, driving under the influence of drugs including marijuana is prohibited in all states.

Table 1
List of auxiliary variables used in the Bayesian multilevel imputation model.

Auxiliary variables	Levels
Driver characteristics	
Age	< 20, 20–29, 30–39, 40–49, 50–59, 60+ years
Sex	male, female
Seatbelt use	yes, no
Survival status	fatal, non-fatal
License status	valid license, no valid license
Number of DWI convictions in previous three years	0, 1, 2+
Police reported drug involvement	yes, no
Blood alcohol concentration	(g/mL)
Vehicle characteristics	
Vehicle class	bus, motor homes, light trucks and vans, minivans, miscellaneous vehicles, median and heavy trucks, motorcycles, passenger cars, utility vehicles
Crash circumstances	
Day of week	Monday–Thursday, Friday, Saturday, Sunday
Time of day	6:00–9:59, 10:00–15:59, 16:00–19:59, 20:00–23:59, 0:00–5:59
State-level information	
Census region	Northeast, Midwest, South, West
State	50 states and the District of Columbia
State marijuana laws legalizing medical or recreational use	yes, no

The police reported drug use was missing for a substantial proportion of cases and the meaning of a missing value could vary from case to case. To address this problem, we treated the police reported drug use as a fully observed three-level covariate, with missing regarded as a separate category (Rubin et al., 1998). Proportions of missing data in the other covariates are small with less than 3.5% in each covariate except seatbelt use (11.9%). We imputed the small proportions of missing data in the covariates using the "mice" package in R (R Development Core Team, Vienna, Austria) (Van Buuren and Groothuis-Oudshoorn, 2011).

2.4. Imputation for missing marijuana data

We calculated the percentages of drivers with positive and negative marijuana test results in each of the 50 states and the District of Columbia using the data from the 34.4% of drivers with marijuana test results. Seven states have exceptionally low testing rates or exceptionally low/high marijuana positivity rates, including DE, IA, ME, MD, MS, NC, and OK (Table 2). We excluded these seven states from the dataset used for building imputation models because data for these states are deemed invalid and unreliable.

We labeled the marijuana test result y_i for driver i with 1 for drivers testing positive for marijuana and 0 for drivers testing negative for marijuana. To impute the dichotomous marijuana variable, we built the following multilevel logistic regression model:

$$\begin{aligned}\text{logit}\{\Pr(y_i = 1)\} &= \beta_0 + \beta^{female}\bullet female_i + \beta^{seatbelt}\bullet seatbelt_i \\ &+ \beta^{survival}\bullet survival_i + \beta^{license}\bullet license_i + \beta^{law}\bullet law_i + \beta^{(BAC\geq 0.01)} \\ &\bullet (BAC_i \geq 0.01) + g(BAC_i) + \alpha_i^{age} + \alpha_i^{preIncident} + \alpha_i^{day} \\ &+ \alpha_i^{time} + \alpha_i^{vehicle} + \alpha_i^{policeRep} + \alpha_i^{state} + \alpha_i^{region} \qquad (1)\end{aligned}$$

In model (1), the dichotomous covariates, including driver sex, seatbelt use, survival status, license status, and state marijuana laws, are included as dummy variables and modeled using a vector of fixed slopes β. The BAC variable is included in the model as both a dummy variable with $(BAC_i \geq 0.01) = 1$ when BAC ≥ 0.01 g/dL and 0 otherwise and a smooth $g\,(BAC_i)$ term to model a dose-response association of the