

First Name:

Raydel

Last Name:

Valdes Salgado

Middle Initial:

Email:

rvaldes@jhsph.edu

Department:

Epi

Advisor:

Frances Stillman

Degree:

EdD

Project Title:

Changes in intensity of daily tobacco consumption in the US. An application of multilevel and longitudinal analysis to the Tobacco Use Supplement of the Current Population Survey 1992-2007.

Category:

measurement

Changes in intensity of daily tobacco consumption in the US. An application of multilevel and longitudinal analysis to the Tobacco Use Supplement of the Current Population Survey 1992-2007.

INTRODUCTION Tobacco consumption is the single largest preventable cause of death worldwide and a steady reduction in the amount smoked is a trait of the path to successfully quitting¹. The main goal of this proposal is to estimate the temporal changes of the number of cigarettes smoked per day (CPD) in the US. I am hypothesizing that race-specific estimates have been hidden behind overall estimates which are the standard for smoking research. In this analysis I am interested in investigating the effects of state-specific tobacco control environments on the CPD and in characterizing whether these effects vary by racial/ethnic groups. The relevance of the topic and the complexity of the analysis motivated me to apply for this award.

METHODS We propose an analysis of the Tobacco Use Supplement of the Current Population Survey (1992-2007) to describe racial/ethnic specific changes in CPD over time and to determine what individual-level and state-level factors are associated with reduction in the number of CPD. The characteristics of this survey will allow us to do a repeated cross-sectional analysis, a multilevel analysis, and a longitudinal analysis to address the issue of racial/ethnic-specific change in average CPD. The specific aims are described below.

Specific aim 1: To estimate racial/ethnic-specific changes over time in the average number of CPD among smokers in the US between 1992 and 2007.

Hypothesis 1: In general, there is a downward trend in the amount of CPD smoked in the US; however the rate of such decline is unequal among racial/ethnic groups and is also different by gender. The association of individual-level characteristics such as age, income, and education with the temporal change in the average number of CPD is specific for each gender and racial/ethnic group.

Specific aim 2: To examine variability across states in the race-specific temporal change of CPD.

Hypothesis 2: Race-specific patterns of amounts smoked (CPD) exist, but vary across states. Gender specific patterns within racial/ethnic groups are different across states.

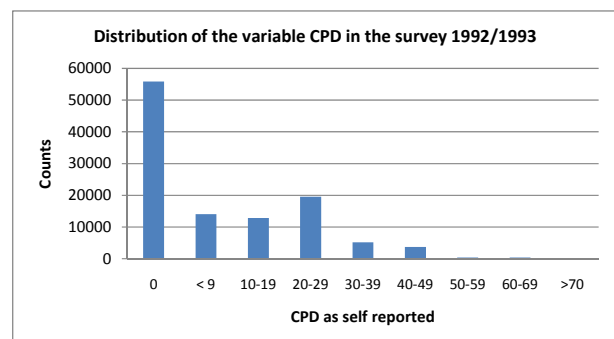
Specific aim 3: By using a longitudinal data set for the period 2002 to 2003, to determine whether increasing cigarette tax affect CPD differently for different racial/ethnic populations after adjusting for other state-level tobacco control measures and individual-level characteristics.

Hypothesis 3: Increasing cigarette tax impacts differently among racial/ethnic groups. Minorities are more responsive to an increase in the price of cigarettes. The impact of such a measure varies across states.

Project' feasibility: The survey includes a probability sample which is allocated among the states to produce state and national estimates. Here, individual-level variables are sociodemographic characteristics and person's smoking behavior, while cigarette price, availability of state-wide secondhand smoke policy at worksites and public places, as well as cessation resources are state-level variables. I gathered information from several sources²⁻⁵ to set the databases needed for this project. Because of the huge sample size (more than 1.3 million of observations) and the complexity of the analysis, to have a powerful computer is a core necessity for this project. Currently, I have not such a machine.

Scientific and scholarly merit: The outcome variable CPD conveys many challenges to the analysis, hence the necessity of effective statistical reasoning and methods to model it. The number of CPD are non-negative, integer-valued responses taking on values (0, 1, 2 ...) which tend to have a mode at zero and a distribution with a long, heavy right tail as shown in the figure. There are two causes of 'real overdispersion' when dealing with count data⁶: one is a violation of the distributional assumptions upon which the model is based; a second cause relates to misspecified variance. With regard to the first cause, we know that the distributional mean of Poisson and NB specifies an expected number of 0 counts in the response, but the

variable CPD has an excess of zeros (happily, most people are non-smokers!⁷).



As for the second cause, we have to be aware of potential extra correlations observed in our models because observations in the surveys are clustered in states. Therefore, we have to use zero-inflated models with state-specific random intercepts for clustered count data to avoid incorrect parameter estimates as well as biased standard errors⁸. In my 30-page proposal I meticulously described the two-part models related to each specific aim and how to interpret their parameters, but here I will only present the random-effects zero-inflated model, with normally-distributed random effects $\varphi_j \sim N(0, \psi_1)$ and $v_j \sim N(0, \psi_2)$ under its general form: $\Pr(y_{ij}) = p(y_{ij}) = \pi_{ij} I(y_{ij}) + (1 - \pi_{ij}) f(y_{ij})$, with

$$\begin{aligned} \text{Part One: } \logit(\pi_{ij}) &= \gamma' w_{ij} + \varphi_j & \text{and} & & \text{Part Two: } \log(\lambda_{ij}) &= \beta' x_{ij} + v_j \\ &= \gamma' w_{ij} + \psi_1 \theta_{1j} & & & &= \beta' x_{ij} + \psi_2 \theta_{2j} \end{aligned}$$

In this approach the probability distribution of y_{ij} is modeled as a function of the covariates w_{ij} and x_{ij} , and the random effects φ_j and v_j are specific to the j^{th} cluster. In other words, the first part involves estimating the probability of a zero outcome (we can say ‘smoking participation’ because it is a binary process, where CPD=0 or CPD>0) and the second part involves estimating the probability of a non-zero count (CPD> 0). The interpretation varies by specific aim, but in general we can say that γ measures the change in the conditional logit of reporting 0 CPD (π_{ij}) for individuals in each cluster j , adjusting by w_{ij} (the individual-level and state-level covariates). Likewise, β measures the change in the conditional log of the number of CPD (λ_{ij}) for individuals in each cluster j , adjusting by x_{ij} (the individual-level and state-level covariates). The random intercepts represent the combined effect of all omitted state-specific covariates that cause individuals from some state to be more prone to smoke than others (φ_j in the Bernoulli part) and to smoke more CPD than others (v_j in the count response model). The parameters ψ_1 and ψ_2 represent the cluster variance terms for the logistic and Poisson components, respectively. Similarly, θ_{1j} and θ_{2j} represent the within cluster variance⁸.

There are additional challenges; first, we have to test that the error terms in Part One are not correlated with error terms of the second part of the models. Additionally, in the longitudinal analysis, we have to consider a serial dependence correlation structure⁹. Because of difference in cluster sample size –particularly for minority groups in some states– the precision of cluster specific estimates may vary greatly. Hence, we will apply Bayesian 2-stage hierarchical models to estimate more efficiently the state-specific and national averages.

SIGNIFICANCE

Why would be relevant for the School to support this project? To the best of our knowledge, there is no precedent of using a national sample for these complex analyses. We are using public data bases, so our study we will be a rich source of examples that can be used in class here at JHSPH. For teaching purposes, this project might illustrate well the challenge due to the hierarchical data collection procedure as zero-inflation and lack of independence may be present simultaneously as a consequence of the inherent correlation structure and underlying heterogeneity. Therefore, our models also account for the unequal racial composition across states and the lack of homogeneity in the strengths of tobacco control programs at state level. These last two features are relevant to any epidemiologic interpretation of our national and state estimates. Very seldom do we have repeated measurements of smoking variables, so a longitudinal analysis is also possible. This kind of statistical analysis almost represents the closest we could be in Epidemiology to a counterfactual scenario to estimate racial/ethnic specific impact of the strongest control measure, increasing cigarette tax.

Significance for Public Health: This research is in a context where tobacco-related health disparities among different racial/ethnic groups have been documented¹⁰ and the question of whether the benefits of tobacco control interventions are reaching all groups has arisen. Thus, we can potentially contribute to a better understanding of racial/ethnic differences in smoking behavior and the issue of whether tobacco control interventions have a differential impact among groups. Additionally, tobacco control efforts vary widely across the US¹¹, so here we have the opportunity to model gender and race specific patterns of use and decrement in CPD in recent past in each of the US states, which is a unique opportunity to understand how tobacco control measures are more or less effective in different scenarios. Thus, results from this study may not only inform policies that may be race/ethnic-specific, but also contribute to define priorities and allocation of resources for prevention in a more effective way. With this study we could generate sound evidence of the linkages between regulatory interventions at different levels and changes in the intensity of smoking across different subgroups of the population.

I have had a lucky learning experience with all the members of my thesis committee: Jonathan Samet and Frances Stillman from the Department of Epidemiology, Francesca Dominici from Biostatistics, and Hugh Waters from Health Policy and Management.

Sincerely, **Raydel Valdés Salgado**

Allocation of expenses

Purchase of laptop.....	\$1,000.00
Total	\$1,000.00

REFERENCES

1. WHO Report on the Global Tobacco Epidemic. The MPOWER package. Geneva, World Health Organization 2008.
2. Orzechowski W and Walker RC: The tax Burden on Tobacco: A historical compilation 2007. Monthly state cigarette tax reports. Volumen 42. Arlington, VA: Orzechowski & Walker
3. American NonSmokers’s Right Fundation at: www.no-smoke.org
4. Campaing for Tobacco-Free Kids at: www.tobaccofreekids.org
5. North American Quitline Consortium at: www.NAQuitline.org
6. Hilbe JM. 2007: Negative Binomial Regression. Cambridge: Cambridge University Press.
7. National Health Interview Survey 2007 (Early release) at: http://www.cdc.gov/nchs/data/nhis/earlyrelease/200709_08.pdf
8. Hur K, Hedeker D, Henderson W, Khuri S, Daley J: Modeling clustered count data with excess zeros in health care outcome research. Health Service & Outcomes Research Methodology 3: 5-20, 2002
9. Lee AH, Wang K: Multilevel zero-inflated Poisson regression modeling of correlated count data with excess zeros. *Statistical Methods in Medical Research* 2006; 15: 47-61.
10. Fagan P, King G, Deirdre L, Petrucci SA, Robinson RG, Banks D, Marable S, Grana R: Eliminating tobacco-related health disparities: Directions for future research. *Am J Public Health*, 2004; 94: 211-217.
11. National Cancer Institute. *Evaluating ASSIST: A Blueprint for Understanding State-level Tobacco Control*. Tobacco Control Monograph No. 17. Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health, National Cancer Institute. NIH Pub. No. 06-6058, October 2006.