Title
Analysis of Multiple Biomarkers Using Structural Equation Modeling

Journal
Tobacco Regulatory Science, 6(4)

ISSN
2333-9748

Authors
Cao, Wenhao
Hecht, Stephen S
Murphy, Sharon E
et al.

Publication Date
2020-07-01

DOI
10.18001/trs.6.4.4

Peer reviewed

HHS Public Access

Author manuscript

Tob Regul Sci. Author manuscript; available in PMC 2022 May 06.

Published in final edited form as:

Tob Regul Sci. 2020 July ; 6(4): 266–278. doi:10.18001/trs.6.4.4.

Analysis of Multiple Biomarkers Using Structural Equation Modeling

Wenhao Cao, **Stephen S. Hecht, PhD**, **Sharon E. Murphy, PhD**, **Haitao Chu, PhD**, **Neal L. Benowitz, PhD**, **Eric C. Donny, PhD**, **Dorothy K. Hatsukami, PhD**, **Xianghua Luo, PhD**

Wenhao Cao, Master of Science Student, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN. Stephen S. Hecht, Professor, Masonic Cancer Center, University of Minnesota, Minneapolis, MN. Sharon E. Murphy, Professor, Masonic Cancer Center, University of Minnesota, Minneapolis, MN. Haitao Chu, Professor, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN. Neal L. Benowitz, Professor, University of California, Department of Medicine, San Francisco, CA. Eric C. Donny, Professor, Wake Forest School of Medicine, Department of Physiology and Pharmacology, Winston-Salem, NC. Dorothy K. Hatsukami, Professor, Masonic Cancer Center and Department of Psychiatry, University of Minnesota, Minneapolis, MN. Xianghua Luo, Associate Professor, Division of Biostatistics School of Public Health and Masonic Cancer Center, University of Minnesota, Minneapolis, MN.

Abstract

Objectives: When examining the relationship between smoking intensity and toxicant exposure biomarkers in an effort to understand the potential risk for smoking-related disease, individual biomarkers may not be strongly associated with smoking intensity because of the inherent variability in biomarkers. Structural equation modeling (SEM) offers a powerful solution by modeling the relationship between smoking intensity and multiple biomarkers through a latent variable.

Methods: Baseline data from a randomized trial (N = 1250) were used to estimate the relationship between smoking intensity and a latent toxicant exposure variable summarizing five volatile organic compound biomarkers. Two variables of smoking intensity were analyzed: the self-report cigarettes smoked per day and total nicotine equivalents in urine. SEM was compared with linear regression with each biomarker analyzed individually or with the sum score of the five biomarkers.

Results: SEM models showed strong relationships between smoking intensity and the latent toxicant exposure variable, and the relationship was stronger than its counterparts in linear regression with each biomarker analyzed separately or with the sum score.

Correspondence Dr. Luo; luox0054@umn.edu.

Conflicts of Interest Statement

Dr. Benowitz serves as a consultant to pharmaceutical companies that market or are developing smoking cessation medications, and has been a paid expert in litigation against tobacco companies. No other potential conflicts of interest were disclosed.

Conclusions: SEM is a powerful multivariate statistical method for studying multiple biomarkers assessing the same class of harmful constituents. This method could be used to evaluate exposure from different combusted tobacco products.

Keywords

biological marker (biomarker); cigarette smoke; latent variable; multivariate statistical method; structural equation modelling

INTRODUCTION

Tobacco smoking continues to be the single leading preventable cause of death and chronic diseases, including cancer, cardiovascular and pulmonary diseases.[1] Tobacco use accounts for about 30% of all cancer deaths in the United States, including about 80% of all lung cancer deaths.[2] Cigarette smoking is the most common type of tobacco use. As recently as 2017, an estimated 14% of U.S. adults (34 million) were current cigarette smokers.[3] Numerous studies have shown a strong dose-response relationship between cigarette smoking and diseases such as lung cancer and heart disease.[4] The self-reported number of cigarettes smoked per day (CPD) was used as the measure of intensity of smoking by many studies.[4]

Smoking-caused disease is a consequence of exposure to toxicants in cigarette smoke.[5] The Food and Drug Administration (FDA) has established a list of harmful and potentially harmful constituents in tobacco products and tobacco smoke, belonging to one or more of the following categories: carcinogens, respiratory toxicants, cardiovascular toxicants, reproductive or developmental toxicants, or additive chemicals and chemical compounds.[6]

Tobacco-related exposure biomarkers have been widely used to distinguish tobacco users from nonusers, demonstrate the effects of smoking cessation or reduction interventions, establish the dose-response relationship with the amount or intensity of use, or demonstrate the impact of tobacco use on potential health outcomes.[7] In this paper, we focus on establishing the dose-response relationship between exposure biomarkers and intensity of smoking. While a number of tobacco exposure biomarkers may be measured in the same study, typically, biomarkers are analyzed separately to determine their individual relationship with variables such as smoking intensity. However, the association between intensity of smoking and an individual biomarker may not be strong enough due to inherent variability in its level, especially when the sample size is small. The lack of statistical power could be related to the fact that one specific toxicant exposure, such as benzene, could come from multiple sources. Even if smoking is the main source of benzene exposure, benzene exposure levels can still be high in non-smokers who live in high traffic areas and near gasoline stations.[8] The lack of statistical power in a single biomarker analysis can be mitigated if more biomarkers assessing the same class of harmful constituents are measured and simultaneously incorporated into a model. Structural equation modeling (SEM), a multivariate statistical technique,[9] offers a potential solution to this problem by using a series of regression equations to model multiple outcomes simultaneously, in our case, multiple biomarkers.

In this secondary data analysis, we aim to analyze the relationship between intensity of smoking and the exposure to a class of toxicants and carcinogens using the baseline data of a recently completed randomized, parallel, double-blind clinical trial by the Center for the Evaluation of Nicotine in Cigarettes (CENIC)[10], a collaborative effort by researchers from 10 U. S. sites led by the University of Pittsburgh and University of Minnesota, for studying the effect of immediate vs. gradual reduction in nicotine content of cigarettes on biomarkers of exposure. Specifically, the baseline data we used include the urinary concentrations of multiple mercapturic acids that are biomarkers of volatile organic chemicals (VOC) present in tobacco smoke: 3-hydroxypropylmercapturic acid (3-HPMA)[11], a metabolite of acrolein; cyanoethylmercapturic acid (CEMA)[11], a metabolite of acrylonitrile; *S*-phenylmercapturic acid (SPMA)[12], a metabolite of benzene; 2-hydroxypropylmercapturic acid (2-HPMA)[13], a metabolite of propylene oxide; and 3-hydroxy-1-methylpropylmercapturic acid (HMPMA)[11], a metabolite of crotonaldehyde and its isomers 2-methylacrolein and methyl vinyl ketone. Two markers of intensity of smoking were analyzed in our study, including the well-studied self-reported CPD as well as a urine biomarker, total nicotine equivalents (TNE), as CPD might lack precision due to faulty recall or rounding errors.[14] TNE is the sum of the urinary concentrations of nicotine, cotinine, 3'-hydroxycotinine and their glucuronides as well as nicotine-*N*-oxide, accounting, for 75–95% of the nicotine dose and is considered the gold standard for daily nicotine intake.[15] Unlike CPD, TNE itself is a biomarker of tobacco exposure, and accommodates individual differences that may influence other tobacco-related exposure biomarkers, and hence is expected to have a stronger dose-response relationship with tobacco exposure biomarkers than CPD.

We proposed to use SEM to estimate the relationship between intensity of smoking (CPD or TNE) and the five cigarette toxicant biomarkers, 3-HPMA, CEMA, HMPMA, 2-HPMA, and SPMA, assuming a latent variable for the overall VOC exposure of a smoker. In addition, we compare the performance of the SEM models with the approach in which each of the five biomarkers is modeled individually using linear regression. We also study the statistical power of different approaches using random samples generated from the CENIC data with different sample sizes.

METHODS

Study Population

The data motivating this study were collected from 1250 participants enrolled from 10 sites in a randomized trial that studied the effect of immediate vs. gradual reduction in nicotine content of cigarettes on smoking-related behavior and biomarkers of smoking exposure. Participants were screened and then completed two weeks of baseline smoking using their own cigarettes, followed by 20 weeks using study product. The study protocol was approved by the University of Minnesota Institutional Review Board. Details of the study design and the participants have been reported.[10] The current study is restricted to the screening and baseline data. Specifically, at the screening visit, all participants reported information on demographic characteristics and at the second baseline visit (ie, the randomization visit) provided first void morning urine for measurement of biomarkers. Between the two baseline

visits, number of cigarettes smoked in the previous day was recorded using an interactive voice response (IVR) system.

Measurements

The baseline exposure variables of interest include 3-HPMA (nmol/mg creatinine), CEMA (nmol/mg creatinine), HMPMA (nmol/mg creatinine), SPMA (pmol/mg creatinine), and 2-HPMA (nmol/mg creatinine), and TNE (nmol/mg creatinine), as described previously. The baseline CPD was calculated as the average of CPDs reported daily via IVR before randomization (up to two weeks before randomization if the two baseline visits were separated more than two weeks). Participants' age, sex, race, and menthol preference were assessed at the screening visit and body mass index (BMI) was measured at the randomization visit.

Statistical analysis

Structural equation modeling is a statistical technique for analyzing complex relationships between multiple variables. The basic concept of SEM is designing a hypothesized model and then using the experimental data to evaluate whether the model assumption is correct. There are two key parts in SEM: *the measurement model* and *the structural model*. The measurement model defines the relationship between measurable variables and non-measurable latent variables, and the structural model delineates the path links and coefficients between a set of variables including latent variables.[16]

As shown in Figure 1 (panel A), we considered the SEM model (Model 1) with five urine VOC biomarkers, 3-HPMA, CEMA, SPMA, 2-HPMA, and HMPMA being associated with one latent variable (referred to as "the VOC exposure"), which was hypothesized to be affected by CPD. Based on a preliminary inspection of the data, these biomarkers approximately had a log-normal distribution, and hence the log transformation was applied on these biomarkers in the subsequent data analysis. We also considered an SEM model (Model 2) adjusting for age, sex, race, BMI, and menthol preference in addition to the primary covariate of smoking intensity, which is depicted in Figure 1, panel B. Model 1 and Model 2 are also referred to as the "unadjusted model" and the "adjusted model", respectively, in this paper.

The SEM models we considered can be described as follows:

$$x_k = \lambda_k y + e_k, k = 1, \ldots, K \quad (1)$$

$$y = \boldsymbol{\gamma}'\boldsymbol{Z} + \varepsilon, \quad (2)$$

where Equation (1) is the measurement model with x_k ($k = 1,\ldots,K$) denoting the observed variables, the five log-transformed biomarkers in our data ($K = 5$), y is the latent variable for the overall VOC exposure, λ_k is the loading of x_k on y, and e_k is the residual; Equation (2) depicts the structural model with $\boldsymbol{Z}$ denoting the variable(s) affecting the latent variable such as CPD in Model 1 and CPD and the additionally adjusted covariates in Model 2, $\boldsymbol{\gamma}$ is the coefficient for $\boldsymbol{Z}$, and ε is the residual. The residual variables, e_k and ε, are assumed

to be independent, zero-mean normal variables. The latent variable was standardized for the ease of comparison between models by constraining the latent variable to have a mean 0 and variance 1.[17] The log-transformed biomarker variables were also standardized by centering it at its mean and dividing it by its standard deviation, so that the loadings of biomarkers of different range can be compared. As a result, the correlation matrix rather than the covariance matrix was analyzed with the maximum likelihood estimation (MLE) method for model estimation. The estimated effect of CPD on the latent exposure variable is interpreted as how many standard deviations change in the latent variable is associated with one unit change in CPD. The loading of each biomarker on the latent variable indicates how many standard deviations increase in the biomarker (in log scale) is associated with one standard deviation increase in the latent variable.

For a comparison purpose, we also performed separate linear regressions to estimate the association of CPD with each biomarker and the sum score of all biomarkers, after log transforming and standardizing the biomarker variable or the sum of biomarkers. The regression coefficient of CPD in the individual linear regressions is then interpreted as how many standard deviations change in the biomarker variable per unit change in CPD, similar to that in the SEM models. In addition, the effects of the other adjusted variables (age etc.) on the latent variable in the SEM models (or the linear regression models) are intercepted similarly as for the CPD. The Z-test statistics (or *z*-values) and p-values are presented for the effect of CPD from all models. Note that a larger *z*-value is an indication of a stronger association between CPD and the latent variable or the biomarker.

In addition, we conducted simulation studies to investigate the small-sample performance of the SEM approach and the linear regression approach using random samples of different sample sizes (N = 50, 100, or 150) drawn repeatedly from the original sample (N = 1250) with a Monte-Carlo size of 1000. The statistical power of the SEM approach, the individual linear regression and the linear regressions with the sum score of all five biomarkers are reported.

We also conducted analyses with TNE instead of CPD as the measure of intensity of smoking. Log transformation was applied on TNE due to the skewed distribution of TNE based on an exploratory analysis. In the SEM models, the estimated effect of standardized log-TNE on the latent exposure variable is interpreted as how many standard deviations change in the latent variable per standard deviation change in log-TNE. The regression coefficient of log-TNE in the linear regressions has similar interpretation as that in the SEM models.

The fit of all studied SEMs were examined by using the comparative fit index (CFI; a CFI ≥ 0.90 was considered a good fit),[18] Root Mean Square Error of Approximation (RMSEA; a RMSEA ≤ 0.06 or ≤ 0.07 was considered a good fit of the model),[19,20] and Standardized Root Mean Square Residual (SRMSR; SRMSR < 0.05 was considered good fit, and ≤ 0.08 deemed acceptable).[18,21] The SEM models were estimated using the R package lavaan.[17] All analyses were performed using R version 3.5.1 (R Core Team), and the R code for the data analysis and simulations are in Supplementary Material.

RESULTS

Several baseline demographic and smoking-related variables of the study population are shown in supplement Table 1 and more can be found in the main paper.[10] On average, the participants smoked 17.1 (standard deviation = 8.6) cigarettes per day at baseline. The average age was 45.5 years, and the average BMI was 29.6 kg/m^2. They were 44% women, 63% white, 29% black, and 8% other race. Menthol cigarettes were smoked by 47% of the smokers.

CPD was strongly associated with the latent variable in both the unadjusted and adjusted SEM models (Table 1). For each unit change in CPD, the VOC exposure latent variable for the unadjusted and adjusted model increased (in the unit of its standard deviation) 0.048 (95% CI, 0.042–0.053) and 0.045 (95% CI, 0.039–0.050), respectively. Compared with the latent VOC exposure variable, the association of each individual biomarker with CPD was weaker (z-value = 17.09 for latent VOC exposure vs. z-values for individual biomarker ranging from 7.08 to 14.15, based on the unadjusted model). The latent VOC exposure's association with CPD was also stronger than the summation of the five VOC biomarkers (z-value = 14.89). Similar results were found with the adjusted model (Model 2), however, with smaller z-values in all models compared with their unadjusted counterparts. In the models with CPD being replaced with TNE (lower panel of Table 1), the VOC exposure for the unadjusted model and adjusted model increased 0.737 standard deviation (95% CI, 0.712–0.761) and 0.705 standard deviation (95% CI, 0.675–0.735) per one standard deviation change in log-TNE, respectively. All TNE models had larger z-values than their CPD model counterparts, showing that TNE would be a better measure of intensity of smoking to use for studying its correlation with VOC exposures. The comparison between VOC exposure and individual biomarkers for the TNE models was similar to that for the CPD models (see Table 1).

The loadings of five biomarkers on the latent variable are reported in Table 2 and shown in Supplement Figures 1 and 2. The unadjusted and adjusted SEM models had very similar results. For CPD, based on the unadjusted model, HMPMA contributed the most to the latent variable (λ = 0.93; 95% CI, 0.89–0.98), followed by 3-HPMA (λ = 0.92; 95% CI, 0.88–0.96), CEMA (λ = 0.82; 95% CI, 0.78–0.87). The least contributions were from 2-HPMA (λ = 0.48; 95% CI, 0.34–0.54) and SPMA (λ = 0.52; 95% CI, 0.47–0.57). The TNE models had similar results as the CPD models. Informed by the 2-HPMA and SPMA results, we performed a more parsimonious SEM model by excluding 2-HPMA or excluding both 2-HPMA and SPMA, the regression coefficients for the new latent variable (0.047 and 0.047 for the CPD and 0.734 and 0.720 for the TNE models, respectively) and the z-value only changed minimally (17.03 and 16.73 for the CPD and 59.22 and 55.73 for the TNE models, respectively).

The effects of the adjusted covariates on the latent variable based on the adjusted SEM models are reported in Table 3. From the SEM model for CPD, older age (γ = 0.01; 95% CI, 0.01–0.02), female sex (γ = 0.28; 95% CI, 0.18–0.38), menthol preference (γ = 0.17; 95% CI, 0.07–0.28), and lower BMI (γ = −0.02; 95% CI, −0.03 to −0.02) were significantly associated with higher VOC exposure, while race was not found significant. Sex (γ = −0.06;

95% CI, −0.03 to 0.14) lost significance (p = 0.19) and BMI became less significant (γ = −0.006; 95% CI, −0.011 to 0.000, p = 0.05) in the SEM model for TNE.

All studied SEMs indicated satisfactory goodness of fit with the CFI ≥ 0.907 and SRMSR ≤ 0.066, except for RMSEA, which was between 0.085 and 0.194 (see Table 2). Table 4 shows the Monte-Carlo simulation results for different random sample sizes. As expected, the SEM had better power than linear regression with either individual biomarker or the sum score of these biomarkers for all different sample sizes. For CPD, when the sample size was small (N = 50), only CEMA and HMPMA's power remained > 0.80 in its individual linear regression model, while the latent VOC exposure variable in the SEM model remained to have a high power (0.89). The power of the linear regression with the sum score of all 5 biomarkers (0.85) was better than each individual linear regression model, but lower than the SEM model. The powers of the TNE models were all better than their CPD model counterparts.

DISCUSSION

While the studied SEMs in this paper were focused on only one latent variable for the VOC biomarkers, obviously, SEM can accommodate more latent variables for different classes of biomarkers. For example, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL), a metabolite of the tobacco specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK),[22] markers of inflammation,[23] oxidative stress,[24] and platelet activation,[25] which have been shown associated with cigarette smoking, could also be analyzed in SEM with multiple latent variables. Depending on researchers' substantive knowledge and the availability of biomarker and other manifest variables, the SEM can be used to effectively model the relationships among different types of biomarkers and smoking behavior variables.

In the analysis of the CENIC data, we found that among the five studied mercapturic acid biomarkers, three (3-HPMA, CEMA, and HMPMA) were highly loaded on the latent variable, while two (2-HPMA and SPMA) were not. Excluding one or both less highly loaded biomarkers did not affect the estimated coefficient or the significance of the effect of the smoking intensity on the latent variable. This shows the robustness of the SEM that we constructed and also suggests that fewer mercapturic acid biomarkers may be collected without sacrificing much of the statistical power of the study. In certain instances, this could save research resources by eliminating the need for multiple mass spectrometric determinations and data analysis. With the simulation studies, we showed that the SEM provided substantial statistical power gains compared with the linear regressions for each individual biomarker or the sum score of biomarkers, which was more obvious when the sample size was small.

In the analysis, TNE showed a stronger association with the VOC exposure than CPD, indicating that TNE is a better measure of intensity of smoking than CPD because per-cigarette smoking behavior and thereby smoke exposure is not taken into account in the latter. It has been shown that the number and size of puffs are key factors that determine per-cigarette smoke exposure, and the amount of nicotine absorbed from a cigarette increased when the total puff volume increased.[26] Toxicants absorbed from a cigarette are also

associated with the puff volume, thus the association between TNE and VOC exposure was stronger than CPD and VOC exposure. Another possible reason could be that both TNE and mercapturic acid biomarkers in the model were creatinine-adjusted. We also found that older age, lower BMI, female sex, and menthol preference were significantly associated with higher VOC exposure, independent of CPD, whereas sex lost its significance in the TNE model. This is possibly because we used creatinine-adjusted TNE in the model, and creatinine was significantly associated with sex in our data (mean [standard deviation] creatinine for women 105 [65] vs. men 141 [76] mg/dL, $p < .001$).

While this study used only the baseline data of the CENIC study, longitudinal data are available in both the CENIC study and many other smoking studies,[27,28] including the ongoing Population Assessment of Tobacco and Health (PATH) study.[29] Future analyses can use longitudinal SEM[30] to analyze the repeatedly measured biomarkers in the CENIC study. Another potential application of the SEM would be in cohort studies to estimate the effect of biomarker exposure on disease risk. The SEM methodology can also be applied for analyzing biomarker data collected from multiple studies by using the so-called “meta-analytic structural equation modeling” or MASEM.[31,32] This methodology can accommodate different numbers of biomarkers from different studies and efficiently synthesize findings across studies.

Note that the SEM we employed assumes a linear relationship between the latent variable and the manifest variables and between the latent variable and variables such as CPD. This assumption could be found restrictive in applications, and to relax the linear assumption, one can apply transformations on the manifest variables or covariates before applying the SEM method as we demonstrated in this paper.

Finally, we want to mention that the use of (multiple) biomarkers as objective measures of smoke exposure, that assumed the self-report and biomarkers are essentially tapping the same construct, is an advantage of our studied SEM models, and that the overall toxicant exposure indexed by the latent exposure variable(s) would be a better measure of disease risk than individual toxicants or self-report intensity of use.

CONCLUSIONS

In this study, we demonstrate how SEM can be used to investigate the association between intensity of smoking and toxicant biomarkers as a more powerful, alternative method to the standard linear regression approach when multiple biomarkers assessing the same class of harmful constituents such as volatile organic compounds are collected, especially for studies with small sample sizes. The SEM methodology demonstrated in this study paper can also help with the estimation of the association of other smoking behavior variables with biomarkers or the effect of biomarkers on disease risk. It also allows different types of biomarkers to be studied in one model by assuming multiple latent exposure variables.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the CENIC study team for collecting, cleaning, and managing the data used in this study.

References

1. Alberg AJ, Shopland DR, Cummings KM. The 2014 Surgeon General's report: commemorating the 50th Anniversary of the 1964 Report of the Advisory Committee to the US Surgeon General and updating the evidence on the health consequences of cigarette smoking. Am J Epidemiol. 2014;179(4):403–412. [PubMed: 24436362]
2. American Cancer Society. Cancer Prevention & Early Detection Facts & Figures 2017–2018. Atlanta, GA: American Cancer Society; 2018:3–4.
3. Wang TW, Asman K, Gentzke AS, et al. Tobacco product use among adults – United States, 2017. Morbidity and Mortality Weekly Report 2018;67(44):1225. [PubMed: 30408019]
4. US Department of Health and Human Services. The Health Consequences of Smoking—50 years of Progress: A Report of the Surgeon General. 2014. Available at: https://www.hhs.gov/surgeongeneral/reports-and-publications/tobacco/index.html Accessed August 13, 2019.
5. Talhout R, Schulz T, Florek E, et al. Hazardous compounds in tobacco smoke. Int J Environ Res Public Health. 2011;8(2):613–628. [PubMed: 21556207]
6. Center for Tobacco Products/U.S. Food and Drug Administration. Harmful and potentially harmful constituents in tobacco products and tobacco smoke; Established list 2012. Available at: https://www.federalregister.gov/d/2012-7727. Accessed August 13, 2019.
7. Chang CM, Edwards SH, Arab A, et al. Biomarkers of tobacco exposure: summary of an FDA-sponsored public workshop. Cancer Epidemiol Biomarkers Prev. 2016;26(3):291–302. [PubMed: 28151705]
8. Tranfo G, Pigini D, Paci E, et al. , Ancona C. Biomonitoring of urinary benzene metabolite SPMA in the general population in central Italy. Toxics. 2018;6(3):37.
9. Hoyle RH. Structural Equation Modeling: Concepts, Issues, and Applications. SAGE; 1995.
10. Hatsukami DK, Luo X, Jensen JA, et al. Effect of immediate vs gradual reduction in nicotine content of cigarettes on biomarkers of smoke exposure: a randomized clinical trial. JAMA. 2018;320(9):880–891. [PubMed: 30193275]
11. Carmella SG, Chen M, Zarth A, et al. High throughput liquid chromatography–tandem mass spectrometry assay for mercapturic acids of acrolein and crotonaldehyde in cigarette smokers' urine. J Chromatogr B. 2013;935:36–40.
12. Carmella SG, Chen M, Han S, et al. Effects of smoking cessation on eight urinary tobacco carcinogen and toxicant biomarkers. Chem Res Toxicol. 2009;22(4):734–741. [PubMed: 19317515]
13. Zarth AT, Carmella SG, Le CT, et al. Effect of cigarette smoking on urinary 2-hydroxypropylmercapturic acid, a metabolite of propylene oxide. J Chromatogr B. 2014;953:126–131.
14. Joseph AM, Hecht SS, Murphy SE, et al. Relationships between cigarette consumption and biomarkers of tobacco toxin exposure. Cancer Epidemiol Biomarkers Prev. 2005;14(12):2963–2968. [PubMed: 16365017]
15. Scherer G, Engl J, Urban M, et al. Relationship between machine-derived smoke yields and biomarkers in cigarette smokers in Germany. Regul Toxicol Pharmacol. 2007;47(2):171–183. [PubMed: 17034917]
16. Everitt BS, Dunn G. Applied Multivariate Data Analysis, 2nd Ed. London: Hodder Arnold; 2001.
17. Rosseel Y lavaan: An R package for structural equation modeling. J Stat Softw. 2012;48(2):1–36.
18. Hooper D, Coughlan J, Mullen MR. Structural equation modelling: Guidelines for determining model fit. Electronic journal of business research methods. 2008;6(1):53–60.
19. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Modeling. 1999;6(1):1–55.

20. Steiger JH. Understanding the limitations of global fit assessment in structural equation modeling. Pers Individ Dif. 2007;42(5):893–898.
21. Byrne BM. Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming, eBook. New York: Psychology Press; 13 May 2013.
22. Carmella SG, Le K, Upadhyaya P, et al. Analysis of N-and O-glucuronides of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol (NNAL) in human urine. Chem Res Toxicol. 2002;15(4):545–550. [PubMed: 11952341]
23. Yasue H, Hirai N, Mizuno Y, et al. Low-grade inflammation, thrombogenicity, and atherogenic lipid profile in cigarette smokers. Circ J. 2006;70(1):8–13. [PubMed: 16377917]
24. Vassalle C, Petrozzi L, Botto N, et al. Oxidative stress and its association with coronary artery disease and different atherogenic risk factors. J Intern Med. 2004;256(4):308–315. [PubMed: 15367173]
25. FitzGerald GA, Oates JA, Nowak J. Cigarette smoking and hemostatic function. Am Heart J. 1988;115(1):267–271. [PubMed: 3276116]
26. Zacny JP, Stitzer ML, Brown FJ, et al. Human cigarette smoking: effects of puff and inhalation parameters on smoke exposure. J Pharmacol Exp Ther. 1987;240(2):554–564. [PubMed: 3806411]
27. Boden JM, Fergusson DM, Horwood LJ. Cigarette smoking and depression: tests of causal linkages using a longitudinal birth cohort. Br J Psychiatry. 2010;196(6):440–446. [PubMed: 20513853]
28. Grana RA, Popova L, Ling PM. A longitudinal analysis of electronic cigarette use and smoking cessation. JAMA Intern Med. 2014;174(5):812–813. [PubMed: 24664434]
29. United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse, and United States Department of Health and Human Services. Food and Drug Administration. Center for Tobacco Products. Population Assessment of Tobacco and Health (PATH) Study [United States] Public-Use Files. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-09-28. Available at: 10.3886/ICPSR36498.v8. Accessed August 13, 2019.
30. Ferrer E, McArdle J. Alternative structural models for multivariate longitudinal data analysis. Struct Equ Modeling. 2003;10(4):493–524.
31. Cheung MW, Chan W. Meta-analytic structural equation modeling: A two-stage approach. Psychol Methods. 2005;10(1):40–64. [PubMed: 15810868]
32. Cheung MW. Meta-Analysis: A Structural Equation Modeling Approach. John Wiley & Sons; 2015.

IMPLICATIONS FOR TOBACCO REGULATION

SEM is an effective analytic tool which can provide important and reliable conclusions linking multiple tobacco exposure biomarkers to potential health effects of tobacco products. Such tools are critical for the effective health effects assessment and regulation of tobacco products.

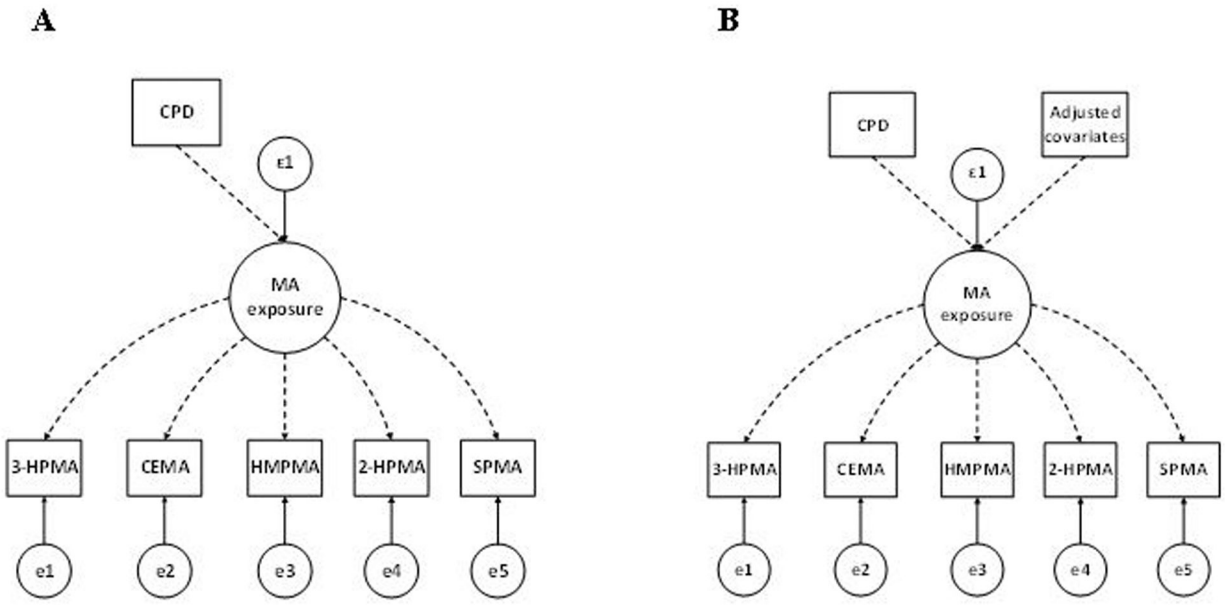

Figure 1. The Structural Equation Modeling for CPD

Note.

A: Unadjusted model (Model 1)

B: Adjusted for covariates of age, gender, race, body mass index, and menthol preference (Model 2).

Large circles represent latent variables; rectangles represent observed or manifest variables; dashed lines with arrows represent the effect from one variable to another; small circles with a solid line pointing to an observed variable represent the residual in the model that cannot be accounted by this observed variable. Note that all the biomarkers in the model were log transformed and standardized. Five observed variables residual were allowed to be correlated in model fitting (not shown in the figure).

Table 1

The Estimated Effects of Intensity of Smoking on Biomarker Exposure Using Structural Equation Models (SEM) and Simple Linear Regression (SLR) Models[a]

	Unadjusted Model		Adjusted Model[b]	
	Regression coefficient (95% CI)	*z*-value[c] (p)	Regression coefficient for CPD (95% CI)	*z*-value[c] (p)
CPD models				
SEM				
Latent VOC exposure	0.048 (0.042, 0.053)	17.09 (< .001)	0.045 (0.039, 0.050)	15.48 (< .001)
SLR with each biomarker				
log(3-HPMA)	0.042 (0.036, 0.049)	13.68 (< .001)	0.040 (0.034, 0.046)	12.71 (< .001)
log(CEMA)	0.043 (0.038, 0.050)	14.15 (< .001)	0.041 (0.035, 0.048)	13.23 (< .001)
log(HMPMA)	0.042 (0.036, 0.049)	13.80 (< .001)	0.039 (0.033, 0.046)	12.68 (< .001)
log(2-HPMA)	0.023 (0.017, 0.029)	7.08 (< .001)	0.021 (0.015, 0.028)	6.31 (< .001)
log(SPMA)	0.034 (0.028, 0.040)	10.73 (< .001)	0.032 (0.026, 0.039)	10.12 (< .001)
SLR with the sum of 5 biomarkers				
log(Total[d])	0.045 (0.039, 0.051)	14.89 (< .001)	0.043 (0.037, 0.049)	14.19 (< .001)
TNE models				
SEM				
Latent VOC exposure	0.737 (0.712, 0.761)	59.84 (< .001)	0.705 (0.675, 0.735)	46.42 (< .001)
SLR with each biomarker				
log(3-HPMA)	0.624 (0.580, 0.667)	28.14 (< .001)	0.601 (0.555, 0.647)	22.54 (< .001)
log(CEMA)	0.719 (0.679, 0.757)	36.41 (< .001)	0.699 (0.659, 0.741)	33.36 (< .001)
log(HMPMA)	0.656 (0.613, 0.697)	30.58 (< .001)	0.622 (0.578, 0.667)	27.47 (< .001)
log(2-HPMA)	0.370 (0.318, 0.421)	14.02 (< .001)	0.338 (0.283, 0.393)	12.07 (< .001)
log(SPMA)	0.539 (0.492, 0.586)	22.55 (< .001)	0.502 (0.453, 0.552)	19.82 (< .001)
SLR with the sum of 5 biomarkers				
log(Total)	0.701 (0.670, 0.749)	35.50 (< .001)	0.668 (0.626, 0.709)	31.65 (< .001)

Note.

a: In all models, natural log transformation was used on five VOC biomarkers (or the sum of biomarkers), TNE. Then in linear regression and SEM they were standardized and in SEM the latent variable was standardized to make the regression coefficients of different model comparable.

b: Adjusted for age, gender, race, body mass index, and menthol preference.

c: *z*-value: the estimated regression coefficient divided by its standard error; larger *z*-values correspond to more significant p-values.

d: Total: add five VOC biomarkers together.

Table 2

Loadings and Model Fit Indices of Unadjusted and Adjusted Structural Equation Models

Variable	Unadjusted SEM Model		Adjusted SEM Model[a]	
	Loading λ (95% CI)	z-value[b] (p)	Loading λ (95% CI)	z-value[b] (p)
CPD models				
log(3-HPMA)	0.92 (0.88, 0.96)	42.03 (<.001)	0.92 (0.88, 0.96)	42.53 (< .001)
log(CEMA)	0.82 (0.78, 0.87)	35.19 (< .001)	0.82 (0.78, 0.87)	35.59 (< .001)
log(HMPMA)	0.93 (0.89, 0.98)	43.14 (< .001)	0.93 (0.89, 0.98)	43.80 (< .001)
log(2-HPMA)	0.48 (0.43, 0.54)	17.69 (< .001)	0.49 (0.43, 0.54)	17.77 (< .001)
log(SPMA)	0.52 (0.47, 0.57)	19.24 (< .001)	0.52 (0.47, 0.58)	19.43 (< .001)
CFI[c]	0.940		0.929	
RMSEA[d]	0.147		0.085	
SRMSR[e]	0.059		0.038	
TNE models				
log(3-HPMA)	0.91 (0.87, 0.95)	45.70 (< .001)	0.91 (0.87, 0.95)	45.97 (< .001)
log(CEMA)	0.84 (0.80, 0.89)	39.48 (< .001)	0.84 (0.80, 0.89)	39.68 (< .001)
log(HMPMA)	0.93 (0.89, 0.96)	47.59 (< .001)	0.93 (0.89, 0.96)	47.99 (< .001)
log(2-HPMA)	0.49 (0.44, 0.54)	18.19 (< .001)	0.49 (0.44, 0.54)	18.19 (< .001)
log(SPMA)	0.54 (0.49, 0.59)	20.66 (< .001)	0.54 (0.49, 0.59)	20.70 (< .001)
CFI[c]	0.913		0.907	
RMSEA[d]	0.194		0.105	
SRMSR[e]	0.066		0.041	

Note.

a: Adjusted for age, gender, race, body mass index, and menthol preference.

b: z-value: the estimated regression coefficient divided by its standard error; larger z-values correspond to more significant p-values.

c: Comparative Fit Index: a CFI ≥ 0.90 was considered a good fit of the model.

d: Root Mean Square Error of Approximation: a RMSEA ≤ 0.06 or ≤ 0.07 was considered a good fit of the model.

e: Standardized Root Mean Square Residual: a SRMSR ≤ 0.05 was considered a good fit and SRMSR ≤ 0.08 was deemed acceptable.

Table 3

The Estimated Effects of the Adjusted Covariates on the MA Exposure Latent Variable

Covariate	CPD model		TNE model	
	γ^a (95% CI)	*z*-value (p)	γ (95% CI)	*z*-value (p)
Age (year)	0.01 (0.01, 0.02)	6.69 (< .01)	0.01 (0.00, 0.01)	4.41 (< .01)
Female	0.28 (0.18, 0.38)	5.36 (< .01)	−0.06 (−0.03, 0.14)	−1.32 (0.19)
Race (reference: white)				
Black	0.02 (−0.19, 0.22)	0.15 (0.88)	0.06 (−0.10, 0.22)	0.76 (0.45)
Other races	0.04 (−0.15, 0.24)	0.45 (0.66)	0.08 (−0.07, 0.24)	1.07 (0.29)
BMI (kg/m^2)	−0.02 (−0.03, −0.02)	−6.13 (< .01)	−0.006 (−0.011, −0.000)	−1.92 (0.05)
Menthol preference	0.17 (0.07, 0.28)	3.19 (< .01)	0.11 (0.02, 0.20)	2.53 (0.01)

Note.

a: γ: standardized effect estimation.

Table 4

Statistical Power of the Unadjusted Structural Equation Model (SEM) and Simple Linear Regression (SLR) Models for Random Samples of Different Sample Sizes

	Power[a]			
	CPD		TNE	
	N[b] = 50	N = 100	N = 50	N = 100
SEM				
Latent MA exposure	0.890	0.989	0.996	0.999
SLR with each biomarker				
log(3-HPMA)	0.780	0.965	0.972	0.999
log(CEMA)	0.812	0.985	0.993	0.999
log(HMPMA)	0.805	0.980	0.979	0.998
log(2-HPMA)	0.269	0.503	0.495	0.678
log(SPMA)	0.548	0.865	0.878	0.963
SLR with the sum of 5 biomarkers				
log(Total)	0.847	0.987	0.985	0.999

Note.

a: Power is the probability of correctly accepting that two variables are related.

b: N: the sample size within each Monte-Carlo simulation.