Title:

Testing Homogeneity in the Time to Onset of Regular Smoking: is there a fast-track subgroup for the trajectory to nicotine dependence?

Short Title:

Testing Homogeneity in the Time to Onset of Regular Smoking

Authors:

Adam Tashman, PhD*
Postdoctoral Researcher
Department of Statistics and Applied Probability
University of California, Santa Barbara
South Hall 5517
Santa Barbara, CA 93106-3110
805-893-5635
tashman@pstat.ucsb.edu

Derek Gordon, PhD
Associate Research Professor
Department of Genetics
Rutgers University
Life Sciences Building 128
Piscataway, NJ 08854
(732) 445-3386
gordon@biology.rutgers.edu

Nancy R. Mendell, PhD
Professor
Department of Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY 11794-3600
(631) 632-8373
nmendell@notes.cc.sunysb.edu

Paichuan Chen
Doctoral Student
Department of Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY 11794-3600
pcchen@ams.sunysb.edu
Laura J Bierut, MD
Professor
Co-Director, Outpatient Psychiatry Clinic
Department of Psychiatry
Washington University School of Medicine
660 South Euclid Avenue, Campus Box: 8134
St. Louis, MO 63110
(314) 362-3492
laura@wustl.edu

Eric O Johnson, PhD
Department of Psychiatry
Henry Ford Health Sciences Center
One Ford Place
Detroit, MI 48202
ejohnso3@hfhs.org

Naomi Breslau, PhD
Professor
Department of Epidemiology
Michigan State University
B601 West Fee Hall
East Lansing, MI 48824
(517) 353-8623 x170
breslau@epi.msu.edu

Dorothy Hatsukami, PhD
Associate Director, Cancer Prevention and Control Research Program
Professor, Department of Psychiatry
Adjunct Professor, Division of Epidemiology and Community Health and
Department of Psychology
Director, Transdisciplinary Tobacco Use Research Center
University of Minnesota
2701 University Ave Se # 201
Minneapolis, MN 55414-3233
(612) 627-1808
hatsu001@umn.edu

Nancy Saccone, PhD
Assistant Professor
Department of Genetics
Washington University School of Medicine
308A Biotechnology Building
St. Louis, MO 63110
(314) 747-3263
Abstract

This paper examines, through mixture analysis, whether smokers are a homogeneous group with respect to the time between smoking a first cigarette and regular smoking. The sampling frame consists of members of a Detroit HMO and drivers from the St. Louis license registry, for a combined sample size of 47,667. There was strong evidence of at least two subpopulations that transition from smoking a first cigarette to the development of regular smoking. Twenty-six percent of subjects progressed to weekly smoking faster than remaining subjects after controlling for race, education, and age at first puff. The participants who made this fast transition were also more likely to become pack-a-day smokers.

Keywords: nicotine dependence, mixture modeling

Acknowledgements

In memory of Theodore Reich, founding Principal Investigator of COGEND, we are indebted to his leadership in the establishment and nurturing of COGEND and acknowledge with great admiration his seminal scientific contributions to the field. We are grateful to Gary Chase for his many contributions to this project and remember with great appreciation his important and diverse scientific accomplishments. Lead investigators directing data collection are Laura Bierut, Naomi Breslau, Dorothy Hatsukami, and Eric Johnson. The authors thank Heidi Kromrei and Tracey Richmond for their assistance in data collection. This work was supported by the NIH grant CA89392 from the National Cancer Institute.
**Financial Disclosures**

Drs. LJ Bierut, AM Goate, AJ Hinrichs, JP Rice, SF Saccone, and JC Wang are listed as inventors on a patent (US 20070258898) held by Perlegen Sciences, Inc., covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. Dr. Bierut has acted as a consultant for Pfizer, Inc. in 2008.
Introduction

Tobacco smoking is widely recognized as a source of severe health consequences including cancer, heart disease, and other major health issues in the smoker [1]. From a biomedical perspective it is useful to consider how variation in susceptibility to nicotine dependence could be instrumental in the prevention and treatment of smoking. There are several robust variables of importance that are associated with smoking behavior and nicotine dependence, such as age, gender, educational attainment, and race. One of the most important predictors of the development of nicotine dependence that has been consistently demonstrated is a young age of onset of smoking [2]. Interestingly, there is also a negative association between age at first cigarette puff and onset of daily smoking, meaning that those who initiate smoking earlier take a longer period of time to transition to regular smoking [3,4]. Though this is initially counterintuitive, it may be explained in part by some of the barriers that adolescent smokers face in obtaining cigarettes.

There is strong evidence that there are genetic predispositions to the development of nicotine dependence. In fact, genetic links to nicotine dependence and the amount of smoking an individual does have been documented [5,6]. Genetic risk factors, once identified, could be most useful in the design of targeted interventions such as prevention and cessation programs.

The aim of this paper is to examine, through mixture analysis, whether smokers are a homogeneous group with respect to the time between smoking a first cigarette and regular smoking. The presence of a mixture of subpopulations would be consistent with potentially different underlying genetic and/or environmental mechanisms that may cause
some subjects to progress faster than others. Such analyses, when combined with genetic studies by our group, offer a powerful method for probing individual and group variation that may be present in nicotine dependence. In addition, we would be able to probe different environmental variables such as parental supervision and peer smoking that may also contribute to different patterns.

A comprehensive explanation of mixture analysis is provided in [7]. Methodologically, researchers have used statistical mixture modeling to study longitudinal smoking patterns [4,8]. Nearly all papers in this stream of research have used discrete mixture modeling applied to longitudinal data [9-11].

The application of mixture modeling to the analysis of cross-sectional smoking data has not been extensive even though this method [12] has yielded a number of findings suggesting that a quantitative trait was in part determined by major genes. For example, [13] applied mixture analysis to body mass index. In [14], mixture analysis is applied to eye-tracking among psychotic patients, normal subjects, and first degree relative. In [15], the approach is applied to intraocular pressure. Additionally, [16] used a mixture approach to detect rare major genes in lipid levels.

In this report, we examine the time between age at first puff and initiation of regular smoking, adjusted for covariates, to determine if this transition represents a homogeneous group. We use a mixture analysis which assumes that each of the given subpopulations follows a normal distribution. We also examine whether smokers with a relatively shorter time to regular smoking (adjusted for covariates) are more likely to become pack-a-day smokers.
Methods

Sample considered

As part of the Collaborative Genetic study of Nicotine Dependence (COGEND), the sampling frame consisted of 34,896 members of a Detroit HMO and 66,641 drivers from the St. Louis license registry. Contact was made with 56,406 respondents, of whom 54,080 were eligible for study participation (that is, spoke English, were between 25 and 44, and were mentally competent). There were 47,667 respondents in the data set after removing individuals with inconsistent data (e.g., respondent’s age at initiation of weekly smoking less than respondent’s age at first puff). The characteristics of this subpopulation are summarized in Table 1.

Table 1 illustrates the definitions used here. A respondent is categorized in state A (that is, was a smoker) if there was an age when first “puffed” in response to the question “How old were you the very first time you ever smoked even a puff of a cigarette?” A respondent is categorized in state B (that is, was a nonsmoker) if the subject never took “even a puff” of a cigarette. Respondents in state B were excluded from further analysis. Respondents in state A were divided into one of two states: state C (smoked at least 100 cigarettes during their lifetime) or state D (total lifetime cigarette usage less than100 cigarettes). Specifically, they were asked “Over your lifetime, have you smoked at least 100 cigarettes?” Respondents in state C were divided into one of two states: state E (became a regular smoker) or state F (not in state E). A
regular smoker (state E) was one who answered yes to the question “Was there ever a period in your life lasting at least two months when you smoked cigarettes at least once a week?” This study focuses on the 14,266 respondents who reached state E and provided complete information on the variables listed below.

[Insert Table 2 about here]

Independent variables

The independent variables were: (1) age at first puff, (2) gender, (3) race indicator variables, (4) city of residence, and (5) education level.

Regression analysis of the time between first puff and regular smoking

The dependent variable was the natural logarithm of the time between first puff and regular smoking. The logarithm was taken to make the residual variance more nearly constant as a function of the fitted values. For subjects with age at first puff equal to age at first time regular smoking, a value of two months was used as the time difference, since the definition of a regular smoker insists that the minimum time to reach this state is two months. Backward elimination was used for variable selection.

The regression residuals - computed as the difference between the dependent variable and the fitting equation - serve as quantitative measures of time between first puff and regular smoking after controlling for the covariates. These residuals will be referred to as the adjusted time to regular smoking. Residuals from this model will include the effects of any other (if any) omitted factors, including genetically-based groups and other environmental factors.
Mixture analysis of regression analysis residuals

Mixture analysis using normal components was applied to the residuals from the regression model based on independent variables whose coefficient was significant at the 0.05 level of significance. The computations were performed using the mclust package in the R programming environment, which is freely available. This software provides maximum likelihood estimates (MLEs) of the mixing proportions, the means of each component, the within-component variance (equal variance was assumed among the components), as well as the log-likelihood of the data assuming a given number of components.

Logistic regression analysis of progression to pack-a-day smoking

A pack-a-day smoker was one who answered yes to the question “Have you ever smoked at least one pack per day for six months or more.” We fit a logistic regression model to test if smokers who initiated regular smoking faster were more likely to become pack-a-day smokers. The independent variables included those variables used in the earlier regression analysis (age at first puff, gender, race indicator variables, city of residence, and education level). Additionally, we used the residuals from the regression as an independent variable.

Results

Of 47,667 respondents, 58.1% ± 0.5% (95% CI) reached state A (had at least one puff). Of those who reached state A (n=27,672), 56.0% ± 0.5% reached state C; that is, they smoked at least 100 cigarettes lifetime. Of those who reached state C (n=15,484), 93.6% ± 0.4% reached state E; that is, they smoked at least once a
week for at least two months.

Regression of time from first puff to regular smoking

The OLS estimates of the regression coefficients for the significant variables are given in Table 3. There was a positive association between highest grade reported as completed and longer time to regular smoking (t=18.00). Subjects who reported themselves as black had a longer time to regular smoking than subjects who reported themselves as white (t=12.05). Additionally, subject who reported themselves as Asian had a somewhat longer time to regular smoking (t=2.47). The model explained 15.6% of the total variation (p<0.001). Figure 1 displays the mean time until initiation of regular smoking for age of first puff between 10 and 22 (sample sizes were limited for subjects with age of first puff outside this range).

Mixture analysis of OLS regression residuals

The fitting equation used to produce the residuals is the following:

\[ \hat{Y} = 5.51 - 2.44 \times \log(AgeFirstPuff) + 0.094 \times HighestGrade + 0.38 \times Black + 0.28 \times Asian \]

The histogram of the residuals (Figure 2) suggests they may be a mixture of two or more normally distributed components. Table 4 presents results of the mixture analysis for the residuals from the regression analysis, considering 1, 2, 3, and 4 component mixtures. The likelihood ratio test statistic (LRT) for two components rather than
homogeneous data equals 2,659 (p-value < 0.0001), indicating that the residuals were not a single-component distribution. The LRT for 3 components rather than 2 and the LRT for 4 components rather than 3 were both significant. For each choice of components (2, 3, or 4), the MLEs obtained indicated a left component with between 24% and 26% of subjects having a mean residual between -1.77 and -1.88. The LRT for 3 components rather than 2 was significant with the right component in the 2 component analysis apparently split into two groups. Additionally, the LRT for 4 components rather than 3 was significant. When the 4 component solution is compared to the 2 component solution, both solutions have a leftmost group with about 25% of smokers. The 4 component solution roughly splits the rightmost component (containing about 75% of smokers) into three additional groups. The relative stability of the identification of the leftmost group suggests that it may have substantive importance.

[Insert Table 4 about here]

[Insert Figure 2 about here]

Logistic regression results

The indicator of whether a subject initiated pack-a-day smoking was regressed on the subjects’ age of first puff, gender, education level, indicators for Asian and Black, and the regression residuals. All of the covariates were significant in this model, shown in Table 5. The estimate of 0.57 for gender was highly significant (p-value < 0.0001) and supports the hypothesis that males are more likely to initiate heavy smoking than females. The estimate of -0.3 for the regression residual was highly significant (p-value < 0.0001), and indicates that a subject with relatively shorter adjusted time to regular smoking is
more likely to become a pack-a-day smoker as compared to a subject with a longer adjusted time to regular smoking.

[Insert Table 5 about here]

Discussion and Conclusions

Age of onset of smoking, including the age of taking a first puff of a cigarette, is an important predictor of the subsequent development of nicotine dependence. This has been demonstrated in numerous studies and is again seen in our data. The regression analysis of the time from first puff to regular smoking is consistent with the findings of [3] of a nonlinear association between age of initiation and time to daily smoking. Though younger age of onset of smoking predicts the development of nicotine dependence, it take a longer period of time for the young smokers to transition to daily smoking, most likely because of the difficulties associated with obtaining cigarettes.

We used mixture analysis of the residuals from the regression model to extend their findings and check for an omitted factor. There was strong evidence of at least two groups that transition from smoking a first cigarette to the development of regular smoking. The “fast” component reached adjusted regular smoking quickly; specifically, 26% of subjects who smoked at least once a week (reached state E) began weekly smoking relatively faster than remaining subjects after controlling for race, education, and age at first puff. These analyses show that participants who made this transition from smoking the first puff to regular smoking relatively quickly were also more likely to become pack-a-day smokers.
These results are another piece of evidence suggesting that there are individuals susceptible to regular smoking quickly after initial exposure. This is consistent with the pattern of findings for smoking [2, 4], alcohol [17], drug use [18,19], cannabis [20,21], cocaine [22,23] and hallucinogens [24]. All report that age at first use is associated with increased risk for dependence. Additionally, while [25] find an association between early onset of drinking and subsequent alcohol dependence in adult twins, they regard the association as non-causal and due to underlying shared liability to earlier age of onset and dependence.

More specifically, [4] estimated prominent features of tobacco dependence in 29% of the recent-onset smokers in a prospective study. Our finding of 26% appears to be consistent with this result, and extends it to pack-a-day smoking. A noteworthy limitation of the mixture approach is that the underlying mechanism determining the mixture component need not be genetic in nature. Nevertheless, detecting homogeneous mixtures within this population of smokers may help elucidate genetic variants that increase risk for smoking, in addition to those already being discovered [6].
Table 1. Characteristics of Regular Smokers in the Collaborative Genetic study of Nicotine Dependence

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of regular smokers</td>
<td>14266</td>
<td>100</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>8149</td>
<td>57.1</td>
</tr>
<tr>
<td>Male</td>
<td>6117</td>
<td>42.9</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American/Black</td>
<td>2787</td>
<td>19.5</td>
</tr>
<tr>
<td>American Indian/Alaska Native</td>
<td>94</td>
<td>0.7</td>
</tr>
<tr>
<td>Asian</td>
<td>145</td>
<td>1.0</td>
</tr>
<tr>
<td>Native Hawaiian/Other Pacific Islander</td>
<td>18</td>
<td>0.1</td>
</tr>
<tr>
<td>White/Caucasian</td>
<td>10687</td>
<td>74.9</td>
</tr>
<tr>
<td>Other (includes Hispanic, Multiracial)</td>
<td>535</td>
<td>3.8</td>
</tr>
<tr>
<td>City of Residence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detroit</td>
<td>4583</td>
<td>32.1</td>
</tr>
<tr>
<td>St. Louis</td>
<td>9683</td>
<td>67.9</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 12 years</td>
<td>4937</td>
<td>34.6</td>
</tr>
<tr>
<td>13-15 years</td>
<td>4418</td>
<td>31.0</td>
</tr>
<tr>
<td>≥16 years</td>
<td>4911</td>
<td>34.4</td>
</tr>
<tr>
<td>Age at First Puff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 10 years</td>
<td>1585</td>
<td>11.1</td>
</tr>
<tr>
<td>11-12 years</td>
<td>2426</td>
<td>17.0</td>
</tr>
<tr>
<td>13-17 years</td>
<td>7994</td>
<td>56.0</td>
</tr>
<tr>
<td>18-21 years</td>
<td>1776</td>
<td>12.4</td>
</tr>
<tr>
<td>&gt;21 years</td>
<td>485</td>
<td>3.4</td>
</tr>
</tbody>
</table>
Table 2. Definition of Smoking States and Estimated Proportions in State

Legend for Table 1. A respondent was categorized in state A if the subject gave an age in response to question “How old were you the very first time you ever smoked even a puff of a cigarette?” A respondent was categorized in state B if the subject never took “even a puff” of a cigarette. A respondent categorized in state C answered yes to the question “Over your lifetime, have you smoked at least 100 cigarettes?” A respondent who answered no was categorized in state D. A respondent in state E was one who answered yes to the question “Was there ever a period in your life lasting at least two months when you smoked cigarettes at least once a week?” A respondent in state F answered no to this question.

* After removing subjects missing covariate information, there were 14,266 subjects reaching state E.
Table 3: Regression Model  
Response: Log(FirstTimeToWeekly) 

\[ n = 14,266 \]

<table>
<thead>
<tr>
<th>Param</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.51</td>
<td>0.14</td>
<td>38.17</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Log(AgeFirstPuff)</td>
<td>-2.44</td>
<td>0.050</td>
<td>-49.54</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>HighestGrade</td>
<td>0.094</td>
<td>0.0052</td>
<td>18.00</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Black</td>
<td>0.38</td>
<td>0.031</td>
<td>12.05</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Asian</td>
<td>0.28</td>
<td>0.11</td>
<td>2.47</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Residual standard error: 1.274 on 14261 degrees of freedom  
Multiple R-Squared: 0.1564,  Adjusted R-squared: 0.1561  
F-statistic: 660.9 on 4 and 14261 DF,  p-value: < 2.2e-16  
Score (logrank) test = 1047 on 7 df,  p=0
Table 4: Normal Component Mixture Analysis of Residuals from Regression

<table>
<thead>
<tr>
<th>Comp</th>
<th>ΔLRT</th>
<th>$\hat{\sigma}^2$</th>
<th>$\hat{p}_1$</th>
<th>$\hat{\mu}_1$</th>
<th>$\hat{p}_2$</th>
<th>$\hat{\mu}_2$</th>
<th>$\hat{p}_3$</th>
<th>$\hat{\mu}_3$</th>
<th>$\hat{p}_4$</th>
<th>$\hat{\mu}_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NA (^1)</td>
<td>1.62</td>
<td>1</td>
<td>$2.2 \times 10^{-4}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>2,659</td>
<td>0.52</td>
<td>0.26</td>
<td>-1.77</td>
<td>0.74</td>
<td>0.63</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>618</td>
<td>0.37</td>
<td>0.26</td>
<td>-1.82</td>
<td>0.68</td>
<td>0.50</td>
<td>0.062</td>
<td>1.98</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>73.8</td>
<td>0.28</td>
<td>0.24</td>
<td>-1.88</td>
<td>0.25</td>
<td>$-2.0 \times 10^{-3}$</td>
<td>0.46</td>
<td>0.76</td>
<td>0.050</td>
<td>2.16</td>
</tr>
</tbody>
</table>

\(^1\)-2logL = -47,405

Legend for Table 3. Estimates assuming one, two, three, and four normally distributed components in the residuals from the regression model are presented. The increase in the likelihood function (ΔLRT) for the specified number of components (compared to one less) are given. All likelihood ratio test statistics are significant at the 0.001 level (1 vs. 2 component, 2 vs. 3, 3 vs. 4). The components are assumed to have equal variance, and the MLE of the variance $\hat{\sigma}^2$ is given. The first component is defined to be the one with smallest mean, and the MLE of the mixing proportion $\hat{p}_1$ and component mean $\hat{\mu}_i$ are given. Components are ordered by the MLE of the component mean.
**Table 5: Logistic Regression Model**  
Response: Ever Pack-A-Day Smoker

\( n = 14,266 \)

<table>
<thead>
<tr>
<th>Param</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.10</td>
<td>0.26</td>
<td>27.00</td>
<td>&lt;2.00 \times 10^{-16}</td>
</tr>
<tr>
<td>Gender (baseline Female)</td>
<td>0.57</td>
<td>0.04</td>
<td>15.38</td>
<td>&lt;2.00 \times 10^{-18}</td>
</tr>
<tr>
<td>Log(AgeFirstPuff)</td>
<td>-1.41</td>
<td>0.086</td>
<td>-16.32</td>
<td>&lt;2.00 \times 10^{-16}</td>
</tr>
<tr>
<td>HighestGrade</td>
<td>-0.23</td>
<td>0.0093</td>
<td>-24.32</td>
<td>&lt;2.00 \times 10^{-16}</td>
</tr>
<tr>
<td>Black</td>
<td>-0.76</td>
<td>0.053</td>
<td>-14.38</td>
<td>&lt;2.00 \times 10^{-16}</td>
</tr>
<tr>
<td>Asian</td>
<td>-0.80</td>
<td>0.19</td>
<td>-4.15</td>
<td>3.29 \times 10^{-5}</td>
</tr>
<tr>
<td>Regression Residual</td>
<td>-0.31</td>
<td>0.015</td>
<td>-20.86</td>
<td>&lt;2.00 \times 10^{-16}</td>
</tr>
</tbody>
</table>

Null deviance: 19496 on 14265 degrees of freedom  
Residual deviance: 17662 on 14259 degrees of freedom
**Figure 1:** Mean Time between First Puff and Regular Smoking by Age at First Puff

**Legend for Figure 1.**
Expected Age = Age at First Puff + Mean Time to Regular Smoking
Figure 2: Histogram of Residuals from Regression Model

Legend for Figure 2. The residuals are the response values log(TimeFirstToWkly) minus the fitting equation, $\hat{Y} = 5.51 - 2.44 \times \log(\text{AgeFirstPuff}) + 0.094 \times \text{HighestGrade} + 0.38 \times \text{Black} + 0.28 \times \text{Asian}$
References


23. Reboussin BA, Anthony JC. Is there epidemiological evidence to support the idea that a cocaine dependence syndrome emerges soon after onset of