Nicotine & Tobacco Research, Volume 13, Number 12 (December 2011) 1266–1275

Original Investigation

Characterizing Patterns of Smoking Initiation in Adolescence: Comparison of Methods for Dealing With Missing Data

Jon Heron, B.Sc., M.Sc., Ph.D.,[1] *Matthew Hickman, B.Sc., M.Sc., Ph.D., FFPH.,*[1] *John Macleod, B.Sc., MBChB., M.Sc., Ph.D.,*[1] *& Marcus R. Munafò, MA., M.Sc., Ph.D.*[2]

[1] *School of Social and Community Medicine, University of Bristol, Bristol, UK*
[2] *School of Experimental Psychology, University of Bristol, Bristol, UK*

Corresponding Author: Jon Heron, B.Sc., M.Sc., Ph.D., *School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK. Telephone: +44-117-3314565; Fax: +44-117-9287325; E-mail: jon.heron@bristol.ac.uk*

Received October 18, 2010; accepted June 29, 2011

Abstract

Introduction: Tobacco use is common and remains one of the leading causes of preventable death in developed countries. Smoking commonly begins in adolescence, and hence, it is important to understand how smoking behavior develops during this period.

Methods: In a U.K.-based birth cohort, we analyzed repeated measures of smoking frequency in a sample of 7,322 young adolescents. Latent class analysis was used to summarize the data, and the resulting classes of behavior were related to a range of smoking risk factors. Results from a complete case analysis were compared with estimation using full-information maximum likelihood (FIML) and estimation using multiple imputation (MI).

Results: Fifty-three percent of the sample reported having smoked a whole cigarette by age 16 years. The longitudinal data were summarized by 4 distinct patterns of smoking initiation: nonsmokers (79.7%), experimenters (10.3%), late-onset regular smokers (5.5%), and early-onset regular smokers (4.5%). Social disadvantage, other substance use, conduct problems, and female sex were strongly related to being a regular smoker; however, no risk factors studied showed any strong or consistent association with experimentation. In the complete case sample, smoking prevalence was lower, and in addition, the association between different smoking patterns and covariates was often inconsistent with those obtained through FIML/MI.

Conclusions: Most young people have experimented with tobacco smoking by age 16 years, and regular smoking is established in a substantial minority characterized by social disadvantage, other substance, use and conduct disorder. Prevention strategies should focus on this subgroup as most children who experiment with tobacco do not progress to regular smoking.

Introduction

The harmful consequences of tobacco use are well established, and it remains one of the leading causes of preventable death in developed countries (Dani & Harris, 2005). One in three adults worldwide use tobacco, with the majority using cigarettes, and while tobacco production and consumption has declined in developed countries over the last thirty years, it has more than doubled in developing countries over the same period (Davis, Wakefield, Amos, & Gupta, 2007). While effective pharmacological and behavioral treatments for smoking cessation now exist, reducing the burden of tobacco-related disease requires the prevention of uptake as well as improvements in methods that facilitate cessation. Smoking commonly begins in adolescence, and about half of those who do not stop smoking will die of a smoking-related disease (Doll, Peto, Boreham, & Sutherland, 2004). Smoking history within an individual can be characterized as consisting of distinct phases including, for example, early experimentation, progression to regular use, development of dependence, and potentially cessation, which in turn is frequently followed by relapse. Despite this general pattern, there is considerable variation within these phases—in particular, many individuals do not progress to regular use and continue as irregular nondependent smokers (sometimes called "chippers").

By age 11 years, about one third of children in the United Kingdom have tried a cigarette, although only 1% smoke every week, and by age 15 years, about two thirds have tried at least one cigarette (Woodhouse, 2004). A recent U.K. survey showed that 20% of 11- to 16-year-olds smoked regularly (Action on Smoking and Health, 2007), with more girls now smoking than boys. This has obvious implications for the future. The majority of adults who are tobacco dependent started smoking as teenagers. In the United Kingdom alone, it is estimated that 3,000 teenagers a week start to smoke (Royal College of Physicians, 1992). Those who start early are more likely to smoke as adults and are less likely to stop (DiFranza et al., 2002; Karp, O'Loughlin, Paradis, Hanley, & Difranza, 2005; Khuder, Dayal, & Mutgi, 1999; Wellman, DiFranza, Savageau, & Dussault, 2004). Despite the overall decline in smoking over the last three decades, in the United Kingdom, cigarette smoking is highest among 20- to 24-year-olds. It is estimated that 38% of males and 35% of

doi: 10.1093/ntr/ntr161
Advance Access published on October 12, 2011

females in this age group are smokers (Office for National Statistics, 2004). Experimentation usually commences between the ages of 11 and 13, and a complex mixture of factors may influence subsequent tobacco use behavior, including biological, attitudinal, interpersonal, and socioeconomic factors. Mental health problems (e.g., depression and anxiety) may also increase the risk of smoking (Patton et al., 1998; Tyas & Pederson, 1998). Dependence often develops rapidly, and it has been suggested that the adolescent brain may be more sensitive to the effects of nicotine (Slotkin, 2002). It is therefore important to understand the patterns and predictors of smoking initiation in adolescence and young adulthood in order to inform the development of more effective prevention.

Longitudinal cohort studies have become a popular source of information on changing patterns of smoking behavior through adolescence. In recent years, a substantial number of studies have modeled such longitudinal data either by using polynomial growth models (Brook, Zhang, Brook, & Finch, 2010; Simons-Morton, 2007; Windle & Windle, 2001) or in combination with a mixture component, for example, Growth Mixture Models (Brook et al., 2010; Colder, Flay, Segawa, & Hedeker, 2008; Orlando, Tucker, Ellickson, & Klein, 2004). As dropout is a common problem with cohort studies, estimation using full-information maximum likelihood (FIML), which allows any participant who responds on one or more occasion to be included in the analysis, is becoming more and more popular. However, this method does not deal with missing data present among covariates, so the sample used for the multivariable analysis can still be depleted. Alternatively, missing data imputation has been used on occasion to address this problem (e.g., Capaldi, Stoolmiller, Kim, & Yoerger, 2009; Duncan, Duncan, Biglan, & Ary, 1998; Guo et al., 2002; Hix-Small, Duncan, Duncan, & Okut, 2004; Li, Duncan, & Hops, 2001); however, we are aware of no substance use papers that have used multiple imputation (MI) prior to the estimation of a mixture model. In the present study, therefore, we aimed to extract distinct patterns of cigarette smoking initiation in a large population-based cohort of adolescents in the United Kingdom and assess their utility against a number of known risk factors for cigarette smoking in adolescence and early adulthood. In addition, we assessed the feasibility of MI within a mixture model setting and compared these results with those obtained using the more traditional approaches.

Methods

Participants

The sample comprised participants from the Avon Longitudinal Study of Parents and Children (ALSPAC; Golding, Pembrey, & Jones, 2001). ALSPAC is an ongoing population-based study investigating a wide range of environmental and other influences on the health and development of children. Pregnant women resident in the former Avon Health Authority (Bristol) in south-west England with an estimated date of delivery between April 1, 1991 and December 31, 1992 were invited to take part, resulting in a "core" cohort of 14,541 pregnancies and 13,973 singletons/twins alive at 12 months of age. The primary source of data collection was via self-completion questionnaires administered at least annually to the mother, her partner, and the study child. Since the age of 7, the cohort has been invited to annual "focus" clinics for a variety of hands-on assessments. More detailed information on the ALSPAC study is available at http://www.alspac.bris.ac.uk. All aspects of the study were reviewed and approved by the ALSPAC Law and Ethics Committee, which is registered as an Institutional Review Board. Approval was also obtained from the National Health Service Local Research Ethics Committees.

Repeated Smoking Measures

The measures of current smoking behavior used in these analyses were collected on three occasions. Current smoking behavior was defined as a four-category ordinal variable with categories "none," "less than weekly" (from here on referred to as *occasional*), "weekly," and "daily" smoking. At 14 and 16 years, the measure was derived by collapsing over levels of a question on current smoking behavior with six response options: "I have only ever tried smoking cigarettes once or twice"/" I used to smoke sometimes but I never smoke cigarettes now"/"I sometimes smoke cigarettes but I smoke less than one a week"/"I usually smoke between one and six cigarettes a week"/"I usually smoke more than six cigarettes a week, but not every day"/"I usually smoke one or more cigarettes every day." At 15 years, a similar variable was derived from information on whether the respondent has "smoked in last 30 days," respondent "smokes weekly," respondent "smokes daily." The 14- and 16-year data were collected via postal questionnaire, while the 15-year data were collected in a clinic setting via a computer terminal. The median ages at data collection were 14 years 2 months, 15 years 5 months, and 16 years 7 months.

Covariates

Covariates considered as risk factors for cigarette smoking in adolescence included (a) demographic variables collected pre-birth around the time of enrollment, which comprised sex, housing tenure (coded as owned/mortgaged, privately rented, subsidized housing rented from council/housing association), crowding status (coded as the ratio of number of residents to number of rooms in house), maternal educational attainment (coded as no high school qualifications, high school, beyond high school), and parity (coded as whether study child is first/second/third child or greater); (b) young person's risky behaviors collected through focus clinic at age 13 years and postal questionnaire at 11 years, which comprised cigarettes use at 13 years (yes/no), alcohol use at 13 years (none/less than weekly/weekly consumption of at least one whole drink), cannabis use at 13 years (yes/no), maximum number of alcohol drinks consumed on one occasion at 13 years (none/up to 4 U/more than 4 U), and conduct problems at 11 years (score of 0–1/2–3/4+ on the conduct problems subscale of the maternal-report Strengths and Difficulties Questionnaire; Goodman & Scott, 1999); and, (c) maternal substance use in the offspring's later childhood collected via questionnaire, which comprised maternal smoking when the young people were 12 years old (yes/no), maternal alcohol consumption also at age 12 years (evidence of bingeing and high weekly consumption derived from detailed record of beers, wines, and spirits consumed in previous week), and maternal cannabis use when the young people were aged 9 years (yes/no).

Statistical Analysis

Latent Class Analysis

We used latent class analysis (LCA) to describe heterogeneity in patterns of response by deriving distinct profiles of smoking behavior. LCA has often been applied in a longitudinal setting (e.g., Croudace, Jarvelin, Wadsworth, & Jones, 2003; Joinson, Heron, Butler, & Croudace, 2009; Munafo, Heron, & Araya, 2008). The aim is to create a latent grouping of the data, which adequately explain the relationship between the observed variables. Starting with a single class, additional classes are added until the various assessments of model fit reach an acceptable level. A number of the statistical criteria (e.g., entropy, Bayesian information criteria, bivariate residuals) were assessed to determine the optimal number of classes—more details in the Supplementary Material. Model fitting was carried out in Mplus version 6 (Muthén & Muthén, 2010) and checked with results obtained with Latent Gold version 4.5 (Vermunt & Magidson, 2005).

Missing Data

The LCA was repeated three times. First, the latent classes were derived for the sample for which all three measures of smoking behavior were available (complete case dataset), that is, obtained through *listwise deletion*. The validity of any results based on this small subset will depend on the degree to which the nonresponse is missing completely at random (MCAR), that is, that nonresponse is neither related to measured nor unmeasured variables. We refer the interested reader to excellent introductory texts on the various types of missing data (MCAR/MAR/MNAR; Graham, 2009; Schafer & Graham, 2002; Sterne et al., 2009).

Second, the latent class estimation was repeated for those with at least one of the three smoking measures present (partially missing dataset). Mplus achieves this through estimation using FIML. Here, the assumption is that missing data are missing at random (MAR) conditional on the repeated measures data that are observed (nonresponse is random, conditional on these data). Rather than imputing data to fill in any of the missing values, FIML directly estimates all parameters using all available data (Enders, 2001; Enders & Bandalos, 2001). As stated in Enders and Bandalos (2001) " . . . under MAR, the partially observed cases provide important information about the underlying marginal distributions of the incomplete variables and hence may reduce the bias that would result from the listwise deletion of cases." Results from regression models involving the output from this latent class model will be referred to and labeled as the "FIML" results to distinguish them from imputation results described below.

The FIML approach deals with missing data among the repeated measures but not within the independent variables. Consequently, as an alternative to FIML estimation, MI was carried out using chained equations (van Buuren, Boshuizen, & Knook, 1999) using the *ice* routine (Royston, 2009) in Stata. This is an iterative procedure, which uses univariable regression equations applied to each variable in turn to predict any missing data, based on the other variables included in the imputation model. Unlike the FIML approach, MI creates multiple datasets over which any imputed data can vary, reflecting the uncertainly in the true values. This approach avoided any drop in sample size, which would otherwise occur when the covariates were regressed on the latent class outcome (see Supplementary Material for more details on the ice procedure). Previous substance use work combining imputation and mixture modeling has simplified the task either by using a single imputed dataset (Hix-Small et al., 2004; Li et al., 2001) or by restricting the imputation step to the covariates (Guo et al., 2002). The derivation of the latent classes was carried out on each imputed dataset in turn; however, the choice of the optimal number of classes was based on the earlier CC/FIML analyses. While Mplus is capable of working with multiply imputed datasets and producing a singled set of pooled results, these pooled results are only useable if the same class ordering is achieved for each imputed dataset. Note that the ordering of classes is somewhat arbitrary, such that different permutations of the same four classes can result from different starting values. In this analysis, it was apparent that the ordering of the classes was not consistent across the datasets (e.g., the nonsmoking classes might be the first class for some imputed datasets and the second class for others), consequently we chose to carry out the LCA on each dataset in turn and pool the results ourselves.

Regression Modeling of Risk Factors

To examine the relationship between risk factors and our latent classes, a two-stage modeling procedure was used. Following the LCA, class assignment probabilities were exported to in Stata version 11-MP2 (StataCorp., 2009). These probabilities were incorporated as an importance weighting (*iweight*) in a series of univariable multinomial logistic regression models. For the imputed data, the 100 imputed datasets were re-stacked and analyzed using by Stata's *mi* routines.

Results

Characteristics of Participants

The starting sample for these analyses is the 13,973 singletons/twins who survived at least until 1 year of age. Of these, 3,038 respondents provided complete data on the three measures (1,245/41% boys and 1,793/59% girls). A further 2,232 missed one response and 2,052 missed two responses, resulting in a dataset of 7,322 with at least one of the three measures (3,351/46% boys and 3,971/54% girls). First, the relationship between level of response (complete/partial/none) and baseline demographics was studied. There was evidence of an association between level of response and gender as well as housing tenure, parity, and maternal education with complete responders being more likely to be female, have parents who own their own home, have a mother educated beyond high school, and have fewer siblings (data available on request). Level of response was also related to smoking frequency. Table 1 shows the relationship between smoking frequency at each time point and whether the respondent provided complete or partial smoking information. There is strong evidence that those who provide partial smoking information are more likely to smoke regularly.

Latent Class Analyses

Nonimputed Data

For the latent class model, there were 64 possible patterns of response. Of these, 46 were observed in the complete case dataset of 3,038 participants. This sample was dominated by a nonsmoking pattern (*none/none/none*), which was the response for

Table 1. Relationship Between Smoking Frequency and Degree of Response

	14-year smoking *n(%)*		15-year smoking *n(%)*		16-year smoking *n(%)*	
	Complete ($n = 3,038$)	Incomplete ($n = 2631$)	Complete ($n = 3,038$)	Incomplete ($n = 2,069$)	Complete ($n = 3,038$)	Incomplete ($n = 2,631$)
None	2,897 (95.4)	2,438 (92.3)	2,626 (86.4)	1,588 (76.8)	2,485 (81.8)	1,413 (77.8)
Occasional	79 (2.6)	67 (2.6)	203 (6.7)	162 (7.8)	225 (7.4)	129 (7.1)
Weekly	24 (0.8)	48 (1.8)	93 (3.1)	100 (4.8)	121 (4.0)	83 (4.6)
Daily	38 (1.3)	78 (3.0)	116 (3.8)	219 (10.6)	207 (6.8)	191 (10.5)
χ^2	$33.22, p < .001$		$112.4, p < .001$		$22.3, p < .001$	

Note. Incomplete samples contain those among the sample of 7,332 who have the measure in question but are missing one of the other two; hence, this value varies across the different time points.

a total of 2,339 (77%) respondents. There was little evidence of regular smoking before the age of 15, with other common response patterns being *none/none/occasional* ($n = 151$, 5.0%), *none/occasional/none* (85, 2.8%), and *none/none/weekly* (52, 1.7%). The partially missing dataset contained an additional 57 response patterns. Based on the model fit criteria displayed in Supplementary Material Table A1, there was good support for the four-class model for both the complete case ($n = 3,038$) and the FIML ($n = 7,322$) models. Model fit statistics for between two- and five-class models can be found in the Supplementary Material along with a more detailed justification of our model choice.

Figure 1 shows the four smoking profiles extracted with the CC and FIML models. These comprise "non-smokers," "experimenters," "late-onset regular smokers," and "early-onset regular smokers." The individual bars indicate the likely behavior of a given class member at each time point. For instance, experimenters have a low probability of reporting smoking at age 14, but by 16, most will report some recent smoking activity, typically at less than weekly frequency. The majority of the respondents (CC: 85.4%; FIML: 80.7%) fall into the nonsmokers group, who have a very low probability of reporting any smoking across the time period. The responses for those reporting greater exposure to smoking were summarized by three latent classes: Early-onset regular smokers (CC: 1.7%; FIML: 3.3%) were mostly daily smokers by age 14 years and all daily smokers by age 16 years; few late-onset regular smokers (CC: 4.3%; FIML: 7.3%) were smoking at 14 years but over 60% were daily smokers by age 16 years; and finally, experimenters (CC: 8.7%; FIML: 8.7%) smoked more commonly on a monthly basis and showed a more gradual increase.

Imputed Data

The average prevalence of the four classes across the 100 imputed datasets was as follows: nonsmokers 79.7% ($SD = 2.2\%$), experimenters 10.3% ($SD = 2.6\%$), late-onset regular smokers 5.5% ($SD = 1.8\%$), and early-onset regular smokers 4.5% ($SD = 1.1\%$), with the earlier FIML results falling within the spread of values obtained from the imputation. These results are as one would expect. Adolescent smoking behavior has previously been shown to be strongly socially patterned in this cohort (Macleod et al., 2008), and in the current manuscript, we report an association between sociodemographic measures and level of response. By incorporating the partial responders through either FIML or MI estimation, we are permitting more of the regular smokers to be included in the analyzed sample and hence obtaining an upwardly revised prevalence of these groups in the ALSPAC cohort. Table 2 shows how the imputed prevalence of smoking behavior varies across response patterns compare with those we were originally able to derive from the observed data. These results show an increasing prevalence of regular users as we move from the complete case (OOO) through moderate missing (OOM/OMO/MOO) and into more severe levels of missing data (OMM/MOM/OMM). The imputed sample of 7,332 contains approximately twice the proportion of daily smokers at each time point.

Covariate Analysis

Table 3 show a series of univariable associations between smoking risk factors and latent class membership resulting from the MI/LCA analysis. These regression estimates were obtained through a series of multinomial logit models (*mlogit*) using the nonsmokers as the reference outcome level. There is a clear and consistent pattern with the majority of the covariates considered being related to the increasingly severe smoking status with increasing strength of association for instance, being female increases your odds of being an experimental smoking by 42%, of being a late-onset regular smoking by 51%, and of being an early-onset regular smoker by 69%. The change in odds is often more marked, for example, maternal smoking at age 12 years increases the odds of being experimental by 42%, while doubling the odds of being a late-onset regular user and quadrupling the odds of being an early-onset regular user. Patterns of association with experimental smoking were similar though in general these effects were weaker and of smaller magnitude. Post-estimation comparisons were then made across the three smoking classes using each in turn as the reference (data not shown). As one would expect given the dose–response nature of many of the associations, there was stronger evidence for differences between experimenters and early-onset regular users than for either of these classes when compared with the late-onset regular users. Of particular interest would be factors that might distinguish between early- and late-onset regular users. Maternal smoking at age 12 years; self-reported smoking at age 13 years; and self-reported weekly alcohol, bingeing, and also cannabis use at age 13 years conferred a risk of being an early- compared with late-onset regular smoker.

Effect of Missing Data Treatment on Conclusions

Table A2 in the Supplementary Material shows the univariable results obtained through CC, FIML, and imputation. Here, log-odds ratios are displayed to permit the use of *SE*s. The first two

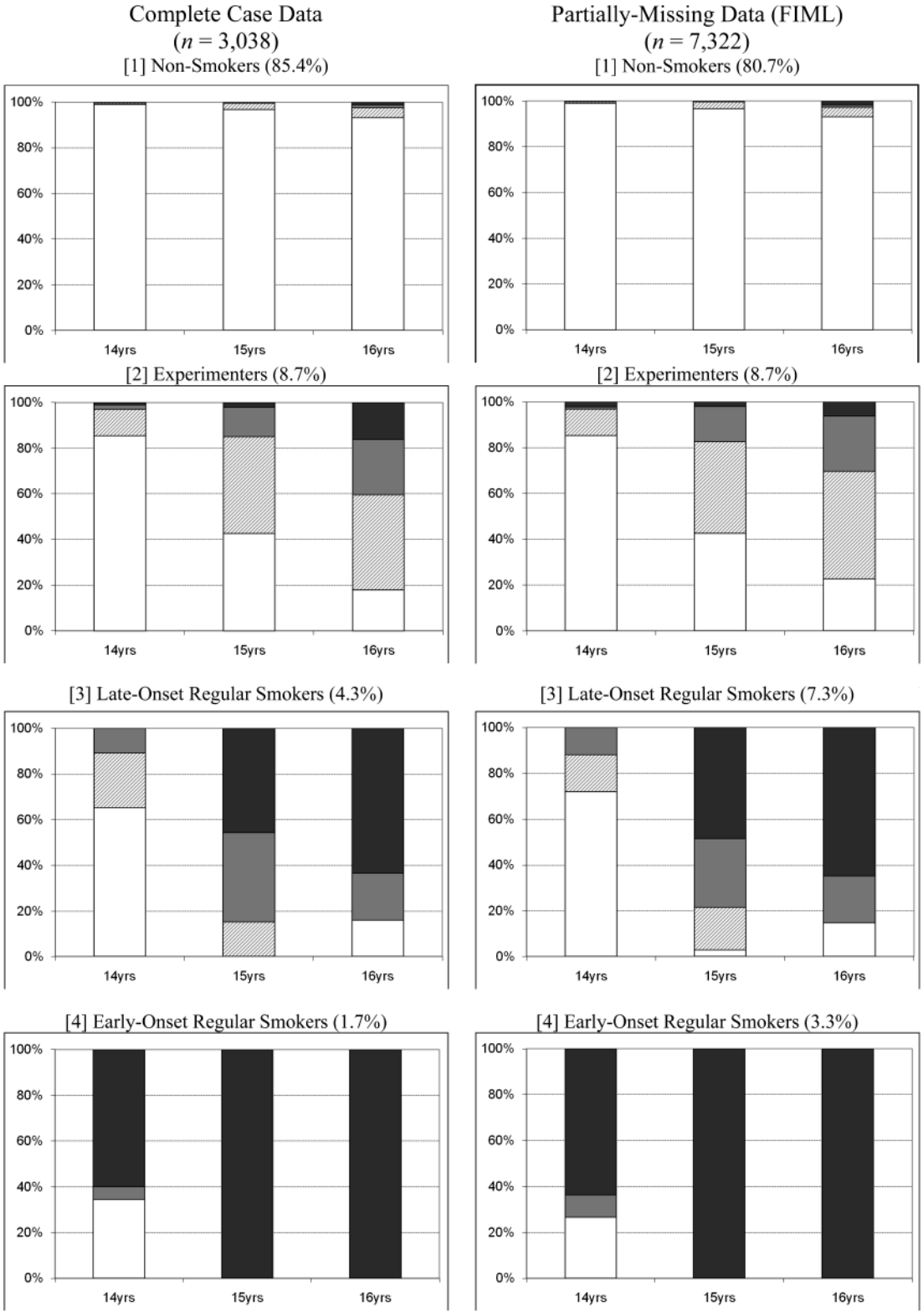

Figure 1. Smoking behavior profiles from four-class model.

Table 2. Observed and Imputed Smoking Frequencies by Missing Data Pattern

	OOO (n = 3,038)	MOO (n = 436)	OMO (n = 990)	OOM (n = 806)	MMO (n = 390)	MOM (n = 827)	OMM (n = 835)	Total (n = 7,332)
14 years, %								
None	95.4	[92.1]	95.2	93.2	[90.5]	[88.6]	89.2	[93.2]
Occasional	2.6	[2.9]	1.8	3.0	[3.4]	[3.9]	3.0	[2.8]
Weekly	0.8	[1.6]	1.2	1.7	[2.0]	[2.4]	2.6	[1.5]
Daily	1.3	[3.4]	1.8	2.1	[4.1]	[5.1]	5.2	[2.6]
15 years, %								
None	86.4	81.7	[83.1]	77.8	[78.5]	73.2	[78.3]	[81.9]
Occasional	6.7	6.7	[7.1]	8.2	[7.3]	8.1	[7.4]	[7.2]
Weekly	3.1	4.6	[3.7]	6.0	[4.4]	3.9	[4.1]	[3.8]
Daily	3.8	7.1	[6.1]	8.1	[9.8]	14.9	[10.3]	[7.1]
16 years, %								
None	81.8	78.0	78.8	[74.6]	75.1	[70.5]	[74.7]	[77.9]
Occasional	7.4	7.1	7.1	[8.3]	7.2	[8.1]	[7.5]	[7.5]
Weekly	4.0	4.4	5.2	[5.0]	3.3	[5.1]	[4.4]	[4.4]
Daily	6.8	10.6	9.0	[12.2]	14.4	[16.3]	[13.4]	[10.2]

Note. O = observed; M = missing; hence OOO indicates the complete cases, MOO indicates those who only missed the 14-year question, OMO missed the 15-year question, etc. Results coming wholly or in part from imputation are shown in square brackets.

columns show a steady drop in sample size as the time since initial enrollment increases. In a univariable analyses, typically 10% of the 3,038 complete cases and 25%–20% of the 7,322 FIML cases will be dropped from the analysis, increasing to approximately a third (CC) and half (FIML) of the observations in any multivariable analysis.

Table 3. Univariable Associations Between Covariates and Latent Class Membership (results for imputed sample, n = 7,322)

		OR (95% CI)			
Covariate	Risk category	Experimenters	Late-onset regular use	Early-onset regular use	p Value
Sex[1]	Female	1.42 (1.17, 1.73)	1.51 (1.18, 1.92)	1.69 (1.29, 2.22)	<.001
Housing tenure[2]	Rented	1.15 (0.83, 1.59)	1.37 (0.92, 2.03)	1.72 (1.12, 2.64)	<.001
	Subsidized housing	1.13 (0.81, 1.59)	1.86 (1.24, 2.80)	3.30 (2.35, 4.63)	
Parity[3]	Second child	1.19 (0.98, 1.45)	1.35 (1.04, 1.75)	1.43 (1.06, 1.92)	<.001
	Third child or higher	1.22 (0.95, 1.57)	1.55 (1.11, 2.16)	2.29 (1.61, 3.25)	
Overcrowding[4]	>1 person/room	1.29 (0.82, 2.02)	2.08 (1.22, 3.57)	3.51 (2.31, 5.35)	<.001
Maternal education[5]	High school qualifications	1.01 (0.82, 1.24)	1.43 (1.01, 2.01)	1.84 (1.34, 2.52)	<.001
	No high school qualifications	1.00 (0.78, 1.29)	1.68 (1.14, 2.48)	2.71 (1.94, 3.79)	
Maternal weekly alcohol at 12 years[6]	14+ U/week	1.23 (1.01, 1.49)	1.17 (0.90, 1.53)	1.21 (0.90, 1.63)	.099
Maternal alcohol binge at 12 years[a]	Yes	1.19 (0.98, 1.43)	1.25 (0.96, 1.63)	1.36 (1.02, 1.80)	.030
Maternal smoking at 12 years	Yes	1.42 (1.09, 1.85)	2.25 (1.61, 3.16)	4.30 (3.07, 6.01)	<.001
Maternal cannabis use at 9 years	Yes	1.93 (1.34, 2.77)	2.57 (1.63, 4.07)	4.07 (2.63, 6.29)	<.001
YP smoking at 13 years	Yes	3.66 (2.59, 5.18)	8.03 (5.23, 12.3)	18.3 (12.1, 27.6)	<.001
YP alcohol at 13 years	Less than weekly	1.82 (1.38, 2.40)	2.26 (1.60, 3.18)	2.88 (1.90, 4.36)	<.001
	Weekly	3.00 (2.16, 4.16)	5.43 (3.61, 8.15)	11.6 (7.70, 17.5)	
YP maximum number of drinks at 13 years	1–4	2.02 (1.58, 2.59)	2.83 (2.12, 3.77)	4.10 (2.86, 5.86)	<.001
	5 or more	3.39 (2.17, 5.30)	7.47 (4.33, 12.9)	19.2 (11.8, 31.4)	
YP cannabis at 13 years	Yes	2.50 (1.33, 4.70)	4.43 (2.10, 9.35)	10.7 (5.91, 19.3)	<.001
YP conduct problems at 11 years[7]	Med	1.23 (0.99, 1.53)	1.80 (1.35, 2.39)	2.20 (1.62, 2.99)	<.001
	High	1.51 (1.01, 2.24)	3.50 (2.21, 5.54)	5.22 (3.44, 7.92)	

Note. Reference categories indicated by superscript: 1: male; 2: mortgaged/owned home; 3: first child; 4: up to one person per room; 5: qualifications beyond high school; 6: weekly use of < 14 U; 7: Score of 0 or 1 on conduct problems (strengths and difficulties); in all other instances, reference = No/None. Ages shown refer to the age of the young person at the time of data collection.

[a]Binge defined as 4+ Units of alcohol on one occasion. One unit of alcohol is equivalent to 0.8 g ethanol.

When performing data imputation, it is of interest to examine the relative contribution to the *SE*s of within- and between-imputation dataset variability. In the current models, we find that, as one might expect, for sociodemographic measures suffering from little nonresponse, the majority (~75%) of the variance comes from *within* each dataset. However, for self-reported substance use measures and other predictors, which suffer from greater levels of nonresponse, the between-dataset contribution has increased considerably to be typically 50% of the total variance (details available on request). The effect of this can be seen in Supplementary Table A2 when comparing the FIML and imputation *SE*s. *SE*s for imputation estimates are *higher* than FIML for covariates that suffer from little missing data (e.g., gender/demographics) since in this situation, the main source of variability is the different latent class estimates across the multiply imputed data; however, as one moves further down the table to covariates more badly affected by dropout, the *SE*s for imputation and FIML become more comparable. Any benefit one might expect from maintaining the sample size at 7,332 using the imputation method is offset by the variability in both the covariate and the outcome across the imputed datasets.

Finally, to summarize the findings as a whole, the FIML and imputation results are broadly consistent with each usually being within one *SE* of the other; however, the bias is clear when examining the complete case estimates as these are often considerably larger or smaller than the other two.

Discussion

We describe patterns on smoking initiation in sample of adolescents from a large representative birth cohort based on reported current smoking frequency at ages 14–16 years. Missing observations, social position, and smoking were related. Following missing data imputation, the classes comprised nonsmokers (80%), experimenters (10%), late-onset regular smokers (5.5%), and early-onset regular smokers (4.5%). About 53% of our sample had ever smoked a cigarette by the age of 16, indicating that in this instance, "experimenters" means those who irregularly use cigarettes over an extended period without developing a consistent pattern of use (and in particular without their use escalating over time). The latent classes had clearly distinct patterns of smoking, with over 60% of the early-onset class smoking daily by age 14 years and all by age 15 years, none of the late-onset class smoking daily by age 14 years and 50% by age 15 years, and weekly smoking being the commonest level of smoking at ages 15 and 16 years among those in the experimenter class. There was good support for a four-class solution across the fit statistics, and there were clear univariable associations between several important risk factors and being in a smoking class which also were stronger for membership of early-onset smoking. These included being female, having older siblings, living in social housing, low maternal education, maternal substance use, and early exposure by the adolescent to tobacco, alcohol, or cannabis.

Previous work describing applying mixture models to smoking initiation in adolescence has typically reported between three and six classes of smoking behavior (e.g., 3: White, Nagin, Replogle, & Stouthamer-Loeber, 2004; 4: Audrain-McGovern et al., 2004; 5: Brook et al., 2008; 6: Pollard, Tucker, Green, Kennedy, & Go, 2010). It has been argued that the number of classes identified in such analyses may be sensitive to a various aspects of the data and chosen model (Jackson & Sher, 2005, 2006, 2008) and that apparent phenotypic variation within- and between-study samples may partly reflect these methodological differences rather than true differences. In the first U.K.-based study to apply these methods, we show that heterogeneity in the frequency of use of tobacco over time can be summarized by four classes of behavior with response profiles consistent with those reported previously, suggesting that patterns of smoking initiation may be similar across countries such as the United States and United Kingdom. Further evidence for the validity of the groupings we identified was reflected in the risk factors that predicted class membership, which have previously been shown to be associated with smoking status more generally. The class prevalences are also consistent with other U.K. surveys, which suggest that 15% of young people are regular smokers (Office for National Statistics, 2003).

The key strength of this study is the size of the cohort, which has been well characterized for multiple factors and has collected contemporaneous repeated measures of tobacco exposure. However, there are several limitations. First, the data were collected in different ways (two postal surveys and one clinic-based assessment) and used slightly different questions. This is the likely cause of the raised residuals in the model fit assessment (see Supplementary Material), which might have led to the extraction of more classes than may have occurred with a more consistent set of questions. Nevertheless, as a data reduction technique, this analysis provided results with good face validity and which performed well against a number of known risk factors for adolescent smoking. Second, our missing data modeling focused on those subjects who had at least one smoking measure from age 14–16 years with the assumption that data were MAR conditional on the range of variables included in the imputation model. This resulted in a sample of 7,322 subjects (52% of the total ALSPAC sample), as opposed to 3,038 (22%) with complete measures. However, as we have demonstrated that smoking behavior and missingness are socially patterned, it is likely that adolescent smoking would be even higher in participants without *any* measures and that the proportion in the smoking classes may be still underestimated compared with the sample of all who originally enrolled in ALSPAC. In an earlier publication based on data from this cohort, Macleod et al. (2008) reported that associations with substance use at age 10 years derived from a complete case analysis were consistent with those from an imputation analysis, despite the difference in substance use prevalence between samples. Our results are consistent with these findings in that missing data techniques mainly led to an increase in apparent smoking prevalence; however, patterns of association between different smoking phenotypes and risk factors were very similar. In the current manuscript, missing data can have an impact in two areas—the latent class derivation and the covariate analysis that follows. Estimation of latent classes under FIML assumes that nonresponse is MAR, that is, the measures we *have* observed are a good representation of each respondent's status across the full time period. There were differences in class size between these results and those obtained from the complete case analysis, which makes the stricter assumption that dropout is MCAR. The imputation approach also assumes MAR but this time conditional on a broader range of measures increasing the chance

that this assumption holds. We feel the MAR assumption to be more justifiable on inclusion of these covariates; hence, the similarity between the FIML and imputation results is encouraging. Nevertheless, we cannot rule out the possibility that smoking behavior may be Not Missing At Random (i.e., nonresponse related to the actual underlying value), and methods that could address this remain underdeveloped. This is a potential limitation to these findings. When it comes to the regression analysis, both CC and FIML samples were reduced due to necessary listwise deletion but to a different degree. There was some benefit to having carried out an initial FIML estimation step; however, the sample size was still seen to impact on the magnitude of some results compared with the imputed models.

Estimation using FIML is the favored method for longitudinal mixture models, now commonplace in the adolescent substance use literature. A problem with this approach is that much of the boost to sample size obtained by including partial nonresponders can then be lost when incorporating covariates. MI is now the standard approach for dealing with nonresponse in epidemiological research. We have shown here that it is now feasible to combine imputation with the estimation of a longitudinal mixture model. To the best of our knowledge, this is the first adolescent substance use study to have done this. There were clear benefits to this approach in the current analysis due to the levels of nonresponse within these data and also the strong social patterning of smoking behavior within this cohort. The future application of these methods may afford greater statistical power by allowing the use of a greater proportion of available data within longitudinal datasets. This is likely to be particularly important when considering influences likely to be of small effect (e.g., genetic variation).

Conclusions

By the age of 16, the majority of children in the United Kingdom have experimented with tobacco smoking. Only a minority, characterized by early onset of use, social disadvantage, other substance use, and conduct disorder, have progressed to regular smoking. It is this minority who are likely to experience the most adverse health effects of smoking in the later life course and who are therefore most likely to benefit from targeted prevention efforts.

Supplementary Material

Supplementary Material can be found online at http://www.ntr.oxfordjournals.org

Funding

The U. K. Medical Research Council (grant 74882), the Wellcome Trust (grant 076467), and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors who will serve as guarantors for the contents of this paper. J.H. is supported by the U.K. Medical Research Council (grants G0800612 and G0802736) and the Wellcome Trust (grant 086684).

Declaration of Interests

None declared.

Acknowledgments

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses.

References

Action on Smoking and Health. (2007). *Essential information on young people and smoking*. Retrieved from http://www.ash.org.uk/files/documents/ASH_108.pdf

Audrain-McGovern, J., Rodriguez, D., Tercyak, K. P., Cuevas, J., Rodgers, K., & Patterson, F. (2004). Identifying and characterizing adolescent smoking trajectories. *Cancer Epidemiology, Biomarkers and Prevention*, *13*, 2023–2034. Retrieved from http://cebp.aacrjournals.org/

Brook, D. W., Brook, J. S., Zhang, C., Whiteman, M., Cohen, P., & Finch, S. J. (2008). Developmental trajectories of cigarette smoking from adolescence to the early thirties: Personality and behavioral risk factors. *Nicotine & Tobacco Research*, *10*, 1283–1291. doi:10.1080/14622200802238993

Brook, D. W., Zhang, C., Brook, J. S., & Finch, S. J. (2010). Trajectories of cigarette smoking from adolescence to young adulthood as predictors of obesity in the mid-30s. *Nicotine & Tobacco Research*, *12*, 263–270. doi:10.1093/ntr/ntp202

Capaldi, D. M., Stoolmiller, M., Kim, H. K., & Yoerger, K. (2009). Growth in alcohol use in at-risk adolescent boys: Two-part random effects prediction models. *Drug and Alcohol Dependence*, *105*, 109–117. doi:10.1016/j.drugalcdep.2009.06.013

Colder, C. R., Flay, B. R., Segawa, E., & Hedeker, D. (2008). Trajectories of smoking among freshmen college students with prior smoking history and risk for future smoking: Data from the University Project Tobacco Etiology Research Network (UpTERN) study. *Addiction*, *103*, 1534–1543. doi:10.1111/j.1360-0443.2008.02280.x

Croudace, T. J., Jarvelin, M. R., Wadsworth, M. E., & Jones, P. B. (2003). Developmental typology of trajectories to nighttime bladder control: Epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology*, *157*, 834–842. doi:10.1093/aje/kwg049

Dani, J. A., & Harris, R. A. (2005). Nicotine addiction and comorbidity with alcohol abuse and mental illness. *Nature Neuroscience*, *8*, 1465–1470. doi:10.1038/nn1580

Davis, R. M., Wakefield, M., Amos, A., & Gupta, P. C. (2007). The Hitchhiker's guide to tobacco control: A global assessment of harms, remedies, and controversies. *Annual Review of Public Health*, *28*, 171–194. doi:10.1146/annurev.publhealth.28.021406.144033

DiFranza, J. R., Savageau, J. A., Rigotti, N. A., Fletcher, K., Ockene, J. K., McNeill, A. D., et al. (2002). Development of symptoms of tobacco dependence in youths: 30 month follow up data from the DANDY study. *Tobacco Control*, *11*, 228–235. Retrieved from http://tobaccocontrol.bmj.com/content/current

Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ*, *328*, 1519. Retrieved from http://www.bmj.com/cgi/content/abstract/328/7455/1519

Duncan, S. C., Duncan, T. E., Biglan, A., & Ary, D. (1998). Contributions of the social context to the development of adolescent substance use: A multivariate latent growth modeling approach. *Drug and Alcohol Dependence*, *50*, 57–71. Retrieved from http://www.sciencedirect.com/science/journal/03768716

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*, 713–740. ISI:000171048800001

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 430–457. doi:10.1207/S15328007SEM0803_5

Golding, J., Pembrey, M., & Jones, R. (2001). ALSPAC—The Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatric and Perinatal Epidemiology*, *15*, 74–87. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291365-3016

Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, *27*, 17–24. Retrieved from http://www.springerlink.com/content/0091-0627/

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Guo, J., Chung, I. J., Hill, K. G., Hawkins, J. D., Catalano, R. F., & Abbott, R. D. (2002). Developmental relationships between adolescent substance use and risky sexual behavior in young adulthood. *Journal of Adolescent Health*, *31*, 354–362. doi:10.5555/ajhb.2007.31.6.672

Hix-Small, H., Duncan, T. E., Duncan, S. C., & Okut, H. (2004). A multivariate associative finite growth mixture modeling approach examining adolescent alcohol and marijuana use. *Journal of Psychopathology and Behavioral Assessment*, *26*, 255–270. Retrieved from http://www.springer.com/psychology/journal/10862

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors*, *19*, 339–351. doi:10.1037/0893-164X.19.4.339

Jackson, K. M., & Sher, K. J. (2006). Comparison of longitudinal phenotypes based on number and timing of assessments: A systematic comparison of trajectory approaches II. *Psychology of Addictive Behaviors*, *20*, 373–384. doi:10.1037/0893-164X.20.4.373

Jackson, K. M., & Sher, K. J. (2008). Comparison of longitudinal phenotypes based on alternate heavy drinking cut scores: A systematic comparison of trajectory approaches III. *Psychology of Addictive Behaviors*, *22*, 198–209. doi:10.1037/0893-164X.22.2.198

Joinson, C., Heron, J., Butler, R., & Croudace, T. J. (2009). Development of nighttime bladder control from 4½–9 years: Association with dimensions of parent rated child maturational level, child temperament and maternal psychopathology. *Longitudinal and Life Course Studies*, *1*, 73–94. Retrieved from http://www.journal.longviewuk.com/index.php/llcs

Karp, I., O'Loughlin, J., Paradis, G., Hanley, J., & Difranza, J. (2005). Smoking trajectories of adolescent novice smokers in a longitudinal study of tobacco use. *Annals of Epidemiology*, *15*, 445–452. doi:10.1016/j.annepidem.2004.10.002

Khuder, S. A., Dayal, H. H., & Mutgi, A. B. (1999). Age at smoking onset and its effect on smoking cessation. *Addictive Behaviors*, *24*, 673–677. doi:10.1016/S0306-4603(98)00113-0

Li, F., Duncan, T. E., & Hops, H. (2001). Examining developmental trajectories in adolescent alcohol use using piecewise growth mixture modeling analysis. *Journal of Studies on Alcohol*, *62*, 199–210. Retrieved from http://www.jsad.com/

Macleod, J., Hickman, M., Bowen, E., Alati, R., Tilling, K., & Smith, G. D. (2008). Parental drug use, early adversities, later childhood problems and children's use of tobacco and alcohol at age 10: Birth cohort study. *Addiction*, *103*, 1731–1743. doi:j.1360-0443.2008.02301.x.

Munafo, M. R., Heron, J., & Araya, R. (2008). Smoking patterns during pregnancy and postnatal period and depressive symptoms. *Nicotine & Tobacco Research*, *10*, 1609–1620. doi:10.1080/14622200802412895

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th Ed.). Los Angeles, CA: Author. Retrieved from http://www.statmodel.com/ugexcerpts.shtml

Office for National Statistics. (2003). *Smoking, drinking and drug use among young people in England in 2002*. London: HMSO.

Office for National Statistics. (2004). *Living in Britain: Results from the 2002 General Household Survey*. London: The Stationery Office.

Orlando, M., Tucker, J. S., Ellickson, P. L., & Klein, D. J. (2004). Developmental trajectories of cigarette smoking and their correlates from early adolescence to young adulthood. *Journal of Consulting and Clinical Psychology*, *72*, 400–410. doi:10.1037/0022-006X.72.3.400

Patton, G. C., Carlin, J. B., Coffey, C., Wolfe, R., Hibbert, M., & Bowes, G. (1998). Depression, anxiety, and smoking initiation: A

prospective study over 3 years. *American Journal of Public Health*, *88*, 1518–1522. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/journals/258/

Pollard, M. S., Tucker, J. S., Green, H. D., Kennedy, D., & Go, M. H. (2010). Friendship networks and trajectories of adolescent tobacco use. *Addictive Behaviors*, *35*, 678–685. doi:10.1016/j.addbeh.2010.02.013

Royal College of Physicians. (1992). *Smoking and the young*. London: Royal College of Physicians.

Royston, P. (2009). Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *Stata Journal*, *9*, 466–477. Retrieved from http://www.stata-journal.com/

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi:10.1037//1082-989X.7.2.147.

Simons-Morton, B. (2007). Social influences on adolescent substance use. *American Journal of Health Behavior*, *31*, 672–684. doi:10.5555/ajhb.2007.31.6.672

Slotkin, T. A. (2002). Nicotine and the adolescent brain: Insights from an animal model. *Neurotoxicology and Teratology*, *24*, 369–384. doi:10.1016/S0892-0362(02)00199-X

StataCorp. (2009). *Stata statistical software: Release 11*. College Station, TX: StataCorp. LP.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, *338*, b2393. PM:19564179

Tyas, S. L., & Pederson, L. L. (1998). Psychosocial factors related to adolescent smoking: A critical review of the literature. *Tobacco Control*, *7*, 409–420. Retrieved from http://tobaccocontrol.bmj.com/

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694. Retrieved from http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291097-0258

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 user's guide*. Belmont, MA: Statistical Innovations. Retrieved from http://www.statisticalinnovations.com/products/latentgold_v4.html#manual

Wellman, R. J., DiFranza, J. R., Savageau, J. A., & Dussault, G. F. (2004). Short term patterns of early smoking acquisition. *Tobacco Control*, *13*, 251–257. Retrieved from http://tobaccocontrol.bmj.com/

White, H. R., Nagin, D., Replogle, E., & Stouthamer-Loeber, M. (2004). Racial differences in trajectories of cigarette use. *Drug and Alcohol Dependence*, *76*, 219–227. doi:10.1016/j.drugalcdep.2004.05.004

Windle, M., & Windle, R. C. (2001). Depressive symptoms and cigarette smoking among middle adolescents: Prospective associations and intrapersonal and interpersonal influences. *Journal of Consulting and Clinical Psychology*, *69*, 215–226. doi:10.1037/0022-006X.69.2.215

Woodhouse, K. (2004). Young people and smoking. In L. Crome, H. Ghodse, E. Gilvarry & P. McArdle (Eds.), *Young people and substance misuse*. London: Royal College of Psychiatrists.