Patterns of Birth Cohort–Specific Smoking Histories, 1965–2009

Theodore R. Holford, PhD, David T. Levy, PhD, Lisa A. McKay, MPH, Lauren Clarke, PhD, Ben Racine, MA, Rafael Meza, PhD, Stephanie Land, PhD, Jihyoun Jeon, PhD, Eric J. Feuer, PhD

Appendix A

Smoking History Model

A compartment model that characterizes a typical smoking history is shown in Figure A1, in which a subject begins to smoke at some point after which they may quit. While this over simplifies what can be much more complex in reality, it does provide a useful characterization of the experience for most of the population. Smoking cessation can be especially difficult to characterize because it is often not successful on the first attempt. Hence, we adopted the rule that subjects who report quitting must have done so at least two years before the interview, otherwise their period of observation is regarded as being truncated at the given age at cessation.

We defined the basic quantities of interest conditional on a hypothetical case with no transitions to death. Let a represent age, t period or calendar year, and c cohort or year of birth, and all three temporal components may play a role when constructing the basic parameters affecting smoking history. These temporal indicators are related by $c = t - a$, therefore, when presenting the relationships among the basic model parameters, we can without loss of generality represent them as functions of age and cohort. The smoking initiation probability, $p(a,c)$, is the conditional probability of smoking initiation at age a for cohort c, if not a smoker at a-1, i.e.,

$$p(a,c) = \Pr\left\{\text{Smoker at } a \,\middle|\, \text{Not smoker at } (a-1), c\right\}.$$

It is related to the cumulative proportion of ever smokers at a conditional on remaining alive,

$$\begin{aligned} P_{\mathrm{E}}(a,c) &= 1 - \prod_{i=1}^{a}\left[1 - p(i,c)\right] \\ &= 1 - \left[1 - P_{\mathrm{E}}(a-1,c)\right]\left[1 - p(a,c)\right] \end{aligned} \qquad (1)$$

where $P_{\mathrm{E}}(0,c) = 0$, which is equivalent to the actuarial approach for estimating the survival curve. If smoking did not affect mortality then one would expect equation (1), which is conditional on remaining alive, to also hold in a population followed over time. But, since mortality is affected by smoking, the observed proportion of the population who have ever smoked at a particular age is given by $P_{\mathrm{E}}^{*}(a,c) \le P_{\mathrm{E}}(a,c)$. Initiation probabilities estimated at a particular survey would be similarly affected by differential mortality; and we represented these by $p^{*}(a,c) = p(a,c)/C_p$ where $C_p \ge 1$ is a constant correction factor introduced to adjust for this effect. We assumed that differential mortality among smoking categories had little effect early in life and the impact intensified with age. Cohorts born before 1935 would only have survey data for ages over 30

$\hat{p}^*(a,c)$

$a \geq a_0$ $P_E^*(a,c) = P_E(a,c)$ for $a < a_0$. Initiation probabilities corrected for differential mortality were found by solving

$$P_E(a_0,c) = 1 - \prod_{i=1}^{a_0}\left[1 - C_p p^*(i,c)\right]$$

for C_p, i.e., by matching the cumulative initiation rates to the estimated prevalence at age a_0. We assumed that a_0 was the age at first survey in 1965 or 30, whichever was older.

Smoking cessation was assumed to be a function of age for each cohort. The smoking cessation probability conditional on the subject being alive and currently smoking is

$$q(a,c) = \Pr\left\{\text{Former smoker at } a \,\middle|\, \text{Smoker at } (a-1), c\right\}.$$

We assumed that $q(a,c) = 0$ for $a < 15$ and we estimated it for $15 \leq a \leq 99$. The cumulative proportion of smokers in cohort c who had not ceased smoking by age a is given by

$$Q(a,c) = \prod_{i=15}^{a}\left[1 - q(i,c)\right] \qquad (2)$$

For simplicity, we assumed that this quantity does not depend on the age an individual who started smoking, number of cigarettes per day or other factors that may be related to an individual's success in quitting. Because initiation tends to occur in a fairly narrow age range, variation in age of initiation becomes less of a factor affecting mortality as a cohort gets older. Introducing intensity of smoking into a model for cessation would require detailed lifetime histories of smoking which were not commonly obtained by NHIS, a limitation in the available data.

Current smokers represent ever smokers who have not quit, and given our assumption that this only depends on age for a given cohort, the prevalence is

$$P_C(a,c) = P_E(a,c)Q(a,c)$$

Former smokers are those who have smoked at some point in their lives, but quit before age a, and the proportion of these individuals is

$$\begin{aligned} P_F(a,c) &= P_E(a,c) - P_C(a,c) \\ &= P_E(a,c)\left[1 - Q(a,c)\right] \end{aligned}$$

Finally, the proportion of cohort c who have never smoked is the complement of those who ever smoked,

$$P_N(a,c) = 1 - P_E(a,c)$$

For a given age and cohort, the sets of current, former and never smokers are exhaustive, i.e.,

$$P_{\mathrm{C}}(a,c)+P_{\mathrm{F}}(a,c)+P_{\mathrm{N}}(a,c)=1$$

Estimation of smoking parameters

Data were only obtained for a restricted range of ages, $a\in[\dot{a},\dot{a}]$, and periods, $t\in[\dot{t},\dot{t}]$ so that the earliest cohort would be $\dot{c}=\dot{t}-\dot{a}$ and the latest $\dot{c}=\dot{t}-\dot{a}$. Available data for a given cohort c, would cover an age range that would vary by cohort, i.e., $a\in[\dot{t}-c,\dot{t}-c]$. To fill in smoking history that was not represented in the survey, we represented each temporal effect as a nonparametric function that we applied outside the range of observed data.

To use this simulation model in a larger decision support framework for planning future strategies for controlling diseases affected by cigarette smoking, birth years 1890 to 2035 have been considered in the model. The earliest birth cohort is represented in the 1965 survey by subjects 75 and older. Because survey participants must be at least 18, the latest cohort was born in 1991 and they would have had a very short smoking history up to that point. Initiation generally occurs early in life, which will usually be better represented in the more recent cohorts, but cessation takes place over the lifespan, which is better represented in older ages by earlier cohorts. NHIS surveys have obtained data during different epochs of life, so it was necessary to extrapolate beyond the range of observed data to obtain estimates for the entire experience of a cohort over its lifespan. The status quo is used to project smoking exposure as a reference for the impact of proposed future interventions on smoking, i.e., age, period and cohort factors were fixed for the period from 2009 to 2050.

Cross-sectional estimates of ever smokers

For years covered by surveys, i.e. 1965-2009, participants provided information that could be used to estimate the prevalence of ever smokers by age, a, for the corresponding cohort, c=t-a. Let Yi be 1 if the i-th individual ever smoked and 0 otherwise, where the probability of the response is a function of age and cohort, $P_{\mathrm{E}}(a,c)$. We assume an additive logistic model for Yi, so that

$$\mathrm{logit}\left\{P_{\mathrm{E}}(a,c)\right\}=\beta_0+\beta_a(a)+\beta_c(c)$$

where β_0 is an intercept and $\beta_{\cdot}(\cdot)$ is a function given by a constrained natural spline (15). The model was fitted using PROC GENMOD in SAS® with knots specified as

Age: 40, 50, 60, 70

Cohort: 1910, 1920, 1930, 1940, 1945, 1950, 1955, 1960, 1965, 1970, 1980

We assumed that the cohort effect remained constant for those born after 1979, the most recent cohort that would provide data to a survey regarding smoking history after age 30 in 2009 which was the age used to identify C_p. Values used for subsequent cohorts were set to be identical to those for the 1980 birth cohort.

Smoking Initiation Probability

Unadjusted estimates of annual age-specific smoking initiation probabilities for a given cohort, $\hat{p}^*(a,c)$, were directly derived from NHIS data. For each cohort represented in a survey, we determined the number of subjects who started to smoke, $d(a,c)$, and who had never smoked to that point, $n(a,c)$. These comprised the response data introduced into a linear logistic model in which the temporal factors were nonparametric functions to be estimated. Each NHIS survey represented participants who survived until that time, and because this group would over represent individuals in a cohort who started smoking late or not at all, these cohort-specific initiation probabilities would be biased downward. The correction factor was found by specifying the target value for the estimated cumulative initiation at a specific age, a_0, to be equal to the value estimated from the cross sectional analysis, i.e.,

$$\hat{P}_{\mathrm{E}}(a^*,c) = 1 - \prod_{i=1}^{a^*}\left[1 - \hat{C}_p \hat{p}^*(a,c)\right]$$

and finding $\hat{C}_p$ which satisfies this condition.

To determine the crude initiation probability estimates, an age-period-cohort model was fitted to the tabulated data given number of subjects who start smoking and are at risk of starting at a given age,

$$\mathrm{logit}\left\{p^*(a,c)\right\} = \beta_0 + \beta_a(a) + \beta_t(t) + \beta_c(c)$$

where β_0 is an intercept and $\beta_\cdot(\cdot)$ is given by a constrained natural spline. We were only interested in the fitted values for the initiation probabilities, which were not affected by the well known identifiability problem in age-period-cohort models (15). Knots were specified as:

age: 10, 15, 20, 50, 60

period: 1910, 1920, 1930, 1940, 1950, 1960, 1970, 1980

cohort: 1910, 1920, 1930, 1940, 1945, 1950, 1955, 1960, 1965, 1970, 1980

Estimates were projected for ages 8-99 to the 2035 birth cohort assuming that the period and cohort effects do not change after 1979, i.e., $\beta_t(t) = \beta_t(1979)$ for p>1979, and $\beta_c(c) = \beta_c(1979)$ for c>1979. Age for the target used to determine the correction factor was age in 1965 (year of the first NHIS survey) or 30, whichever was older, $a^* = \max\{1965 - c, 30\}$. The target value for the cumulative probability of being a smoker was the estimate derived in the analysis of the prevalence curve, $\hat{\Pi}(a^*,c)$.

Smoking Cessation Probability

An individual was identified as having quit smoking if they had not smoked for two years. Because of the two-year lag used in the definition of quitting, an individual who reports cessation at age *a-2* or later could not be classified and they would be truncated at that age. Hence, current smokers were similarly truncated at age *a-2*.

Data used for this analysis were from surveys conducted from 1970-2001 including subjects reporting ages from 17-98. If the reported age of cessation was younger than 8, it was set to 8. For each year of age following smoking, a binary response was created based on our definition of quitting. Yearly estimates of the linear logistic age-period-cohort model was fitted in which

$$\text{logit}\left\{q(a,t,c)\right\} = \beta_0 + \beta_a(a) + \beta_t(t) + \beta_c(c)$$

where β_0 is an intercept and $\beta_\cdot(\cdot)$ are given by a constrained natural splines. We were only interested in the fitted values for the cessation probabilities, which are not affected by the well known identifiability problem in age-period-cohort models (15). Knots were specified as follows:

age: 30, 40, 50, 60

period: 1920, 1930, 1940, 1950, 1960, 1970, 1980

cohort: 1910, 1920, 1930, 1940, 1950, 1960, 1970

Estimates of the yearly cessation probability for age *a* and cohort *c* were the fitted values for ages 15-99, $\hat{q}(a,a+c,c)$. The values used to describe a hypothetical future population in the *status quo* scenario were assumed to have an age effect that remained the same, and the period and cohort effects would remain the same as the estimates for 1979, i.e., $\beta_t(t) = \beta_t(1979)$ for *t*>1979, and $\beta_c(c) = \beta_c(1979)$ for c>1979. The conditional cessation probabilities were used to generate the cumulative probabilities of not quitting, $\hat{Q}(a,a+c,c)$, using equation (2).

Cigarettes Smoked Per Day

Reports of the number of cigarettes smoked per day showed an extremely high degree of digit preference, especially concentrated at half or whole US packs. Therefore, dose was analyzed as an ordered categorical response with half pack being at the center of the category, which was also usually the mode and close to the mean. The intervals (approximate interval center) employed were: CPD≤5 (3); 5< CPD≤15 (10); 15<CPD≤25 (20); 25<CPD≤35 (30); 35<CPD≤45 (40); and 45<CPD (60). A cumulative logistic model was fitted to the data using PROC LOGISTIC in SAS® with age, period and cohort represented by additive nonparametric factors function of time using constrained natural splines. Knots were specified as:

Age knots: 25, 30, 35, 40, 45, 50, 55, 60, 65, 70

Period knots: 1970, 1975, 1980, 1985, 2000, 2005

Cohort knots: 1910, 1920, 1930, 1940, 1950, 1960, 1970, 1980

The fitted estimates of the probabilities for each category of smoking dose for each cohort for ages 0 to 99 were used as parameters for the smoking history generator. Estimates for cohorts born before 1920 were constrained to be the same as for the 1920 birth cohort. Similarly, for cohorts born after 2002 were constrained to be identical to those of the 2002 cohort, who would be 7 in 2009, i.e., the year before the earliest age at initiation considered in this analysis.

As a cohort of smokers gets older, their smoking intensity will change. The SHG implemented this by allowing some individuals to change from one smoking intensity level to another using a Markov process in which this transition from age *a-1* to age *a* only depends on the state at *a-1*. At younger ages, intensity tends to increase, so those with lower intensity may switch to a higher one, and for older ages the reverse occurs. Let $R_i(a,c)$ represent the proportion of subjects with smoking intensity level *i* at age *a* in cohort *c*, and $\sum_{j=1}^{N} R_j(a,c) = 1$ where N is the number of intensity categories. The joint probability distribution for categories at ages *a-1* and *a* in cohort *c* is

$$T_{ij}(a,c) = \Pr\{i \text{ at age } (a-1) \text{ and } j \text{ at age } a \,|\, c\}$$

With marginal distributions given by

$$\sum_{j=1}^{N} T_{ij}(a,c) = R_i(a-1,c)$$

$$\sum_{i=1}^{N} T_{ij}(a,c) = R_j(a,c)$$

Given the marginal distribution for smoking intensity at ages *a-1* and *a*, the matrix for the joint probabilities is constructed using a recursive function starting with

$$T_{11}(a,c) = \min\left[R_1(a-1,c), R_1(a,c)\right]$$

Subsequent iterations used

$$T_{ij}(a,c) = \min\left[R_i(a-1,c) - \sum_{1\le j^*<j} T_{ij^*}(a,c), R_i(a,c) - \sum_{1\le i^*<i} T_{i^*j}(a,c)\right]$$

This allocation formula ensured a proper bivariate distribution that satisfied the specified marginal distribution for smoking intensity.

In order to maintain the appropriate intensity distribution over the life of a cohort, at each year of age, subjects were randomly allocated to a category with transition probability

$$S_{ij}(a,c) = \Pr\{j \text{ at } a \,|\, i \text{ at } a-1,\, c\}$$
$$= T_{ij}(a,c) / R_i(a-1,c)$$

which only depends on the intensity at *a-1*. These transition probabilities satisfied $\sum_{j=1}^{N} S_{ij}(a,c) = 1$. Using the approximate center for each smoking intensity category, X_j, the mean smoking intensity at age *a* in cohort *c* was computed by $\mu(a,c) = \sum_{j=1}^{N} X_j R_j(a,c)$.

Estimation of Current, Former, and Never Smokers for 1-Year Cohorts

Estimates of smoking prevalence were derived from the estimated curves for ever smokers, $\hat{P}_{\mathrm{E}}(a,c)$, and the corresponding survival function for not quitting, $\hat{Q}(a,c)$. The estimated prevalence of current smokers by age and cohort is

$$\hat{P}_{\mathrm{C}}(a,c) = \hat{P}_{\mathrm{E}}(a,c)\hat{Q}(a,c).$$

Prevalence of former smokers is

$$\begin{aligned}\hat{P}_{\mathrm{F}}(a,c) &= \hat{P}_{\mathrm{E}}(a,c) - \hat{P}_{\mathrm{C}}(a,c) \\ &= \hat{P}_{\mathrm{E}}(a,c)\left[1 - \hat{Q}(a,c)\right]\end{aligned}$$

Finally, prevalence of never smokers is

$$\hat{P}_{\mathrm{N}}(a,c) = 1 - \hat{P}_{\mathrm{E}}(a,c).$$

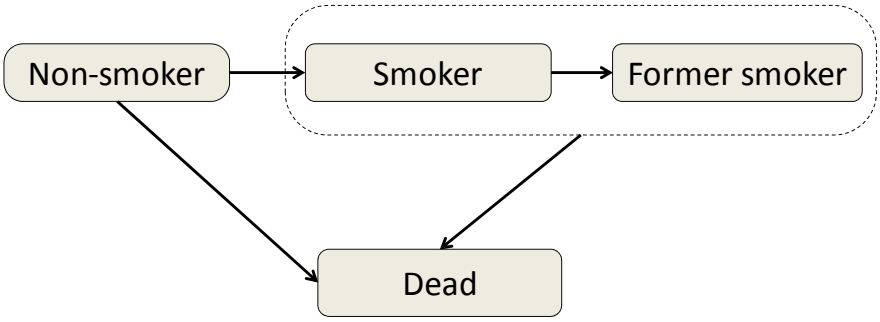

Figure A-1. Compartments considered in developing smoking history.