Birth Cohort-Specific Smoking Patterns by Family Income in the US

Supplementary Material

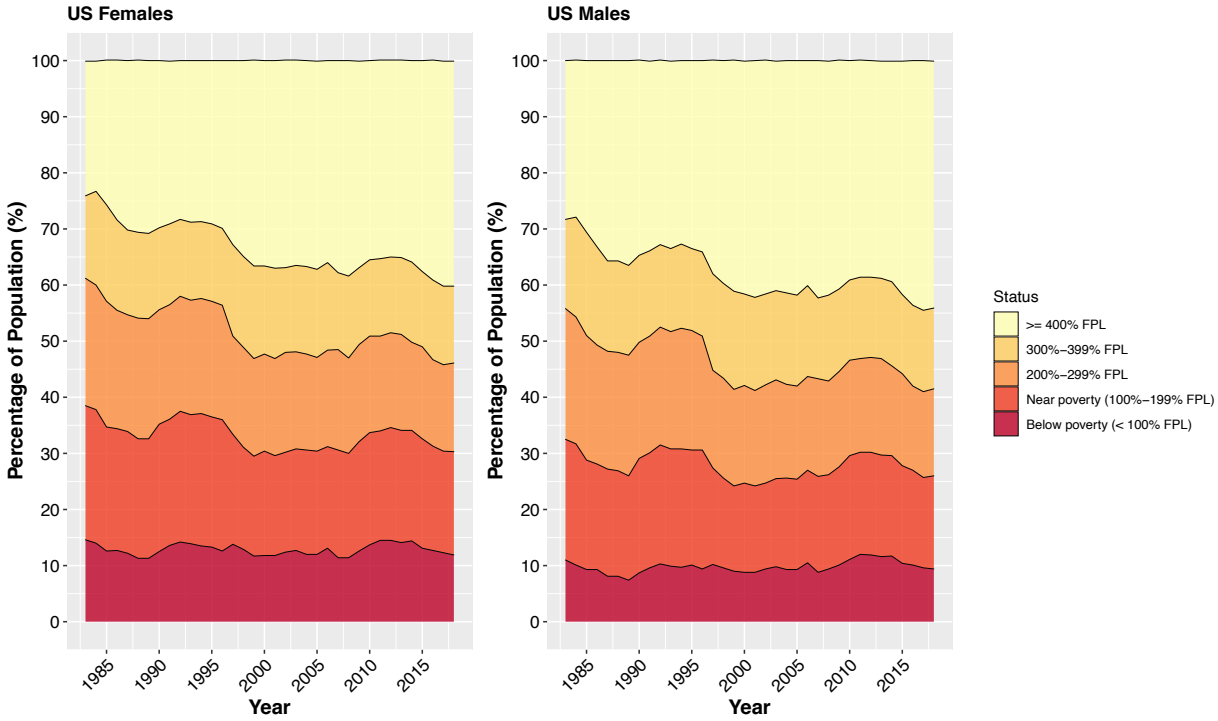

Figure S1. Proportion of US adults aged 18+ (females – left panel, males – right panel) by family income-to-poverty ratio from 1983-2018 in the NHIS data; Below poverty (<100% FPL), Near poverty (100%-199% FPL), 200%-299% FPL, 300%-399% FPL, and ≥400% FPL. FPL=Federal Poverty Level.

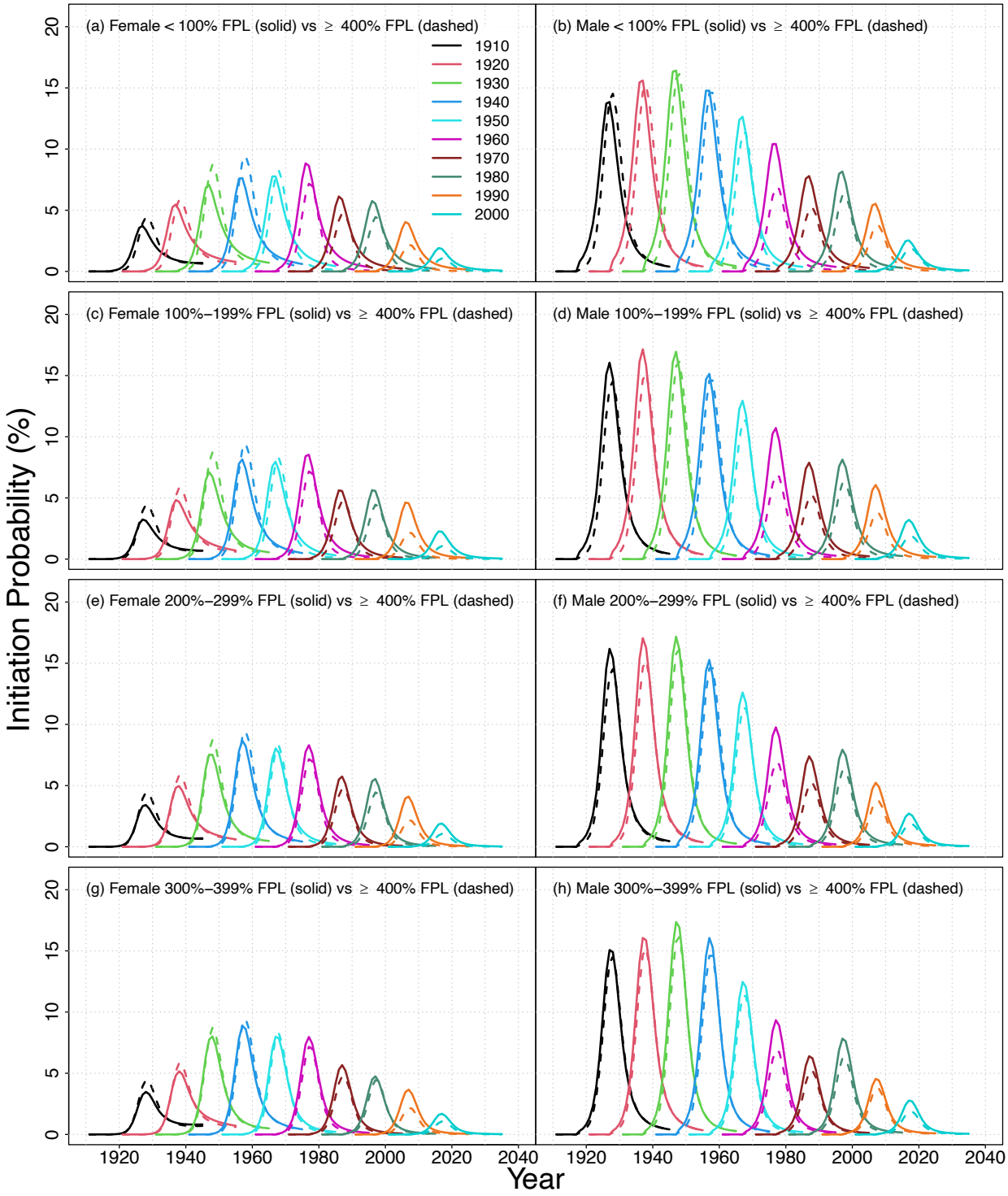

Figure S2. Age-specific smoking initiation probabilities for selected birth cohorts by family income-to-poverty ratio and gender (females – left panels, males – right panels). Solid lines correspond to Below poverty (<100% FPL, panels a & b), Near poverty (100%-199% FPL, panels c & d), 200%-299% FPL (panels e & f), and 300%-399% FPL (panels g and h). Dashed lines represent the initiation probabilities for ≥400% FPL which are shown as reference in all panels. FPL=Federal Poverty Level. An interactive version of this figure's data can be found at: https://sph-umich.shinyapps.io/shgdisplayappincome/

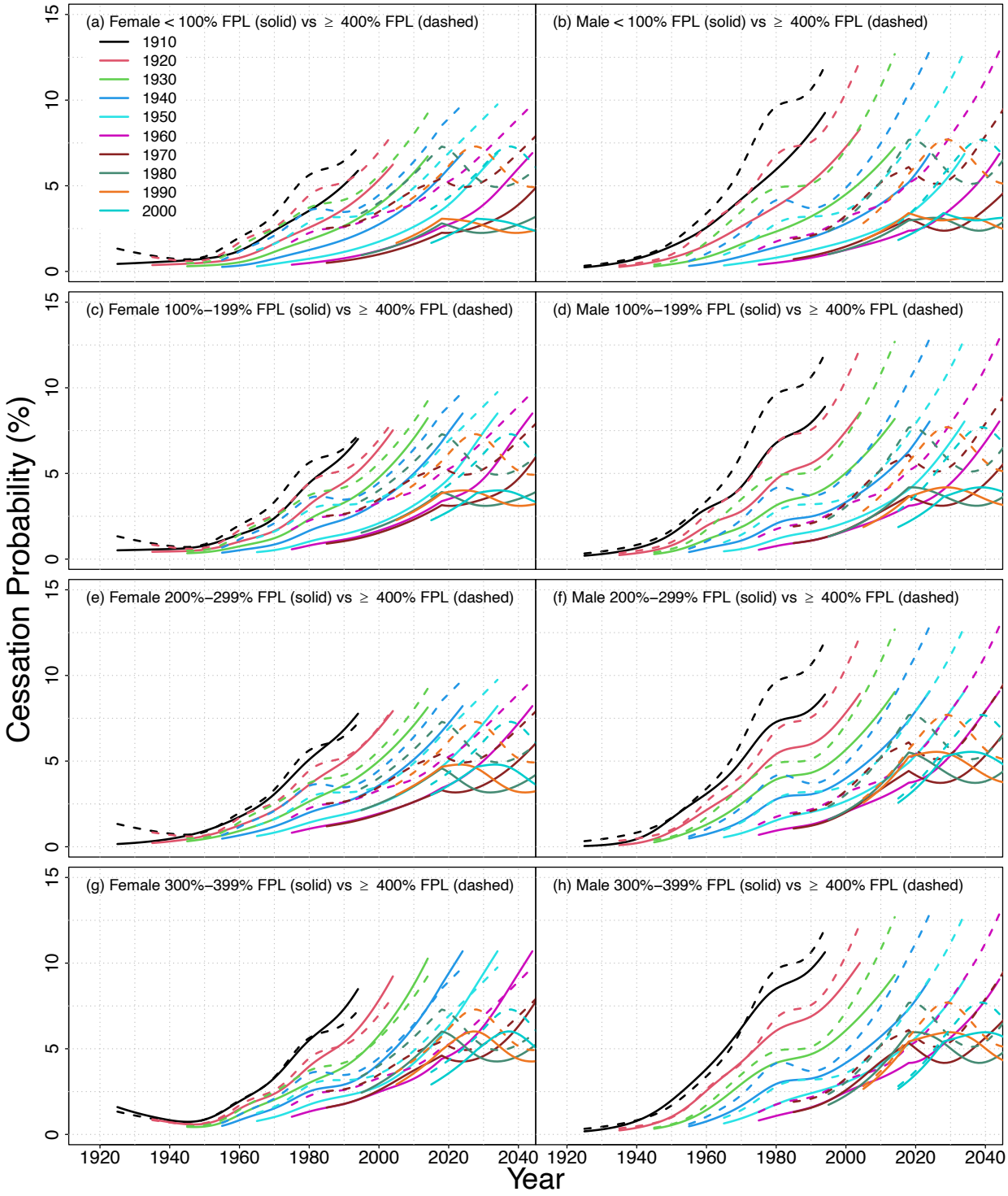

Figure S3. Age-specific smoking cessation probabilities for selected birth cohorts by family income-to-poverty ratio and gender (females – left panels, males – right panels). Solid lines correspond to Below poverty (<100% FPL, panels a & b), Near poverty (100%-199% FPL, panels c & d), 200%-299% FPL (panels e & f), and 300%-399% FPL (panels g and h). Dashed lines represent the cessation probabilities for ≥400% FPL which are shown as reference in all panels. FPL=Federal Poverty Level. An interactive version of this figure's data can be found at: https://sph-umich.shinyapps.io/shgdisplayappincome/

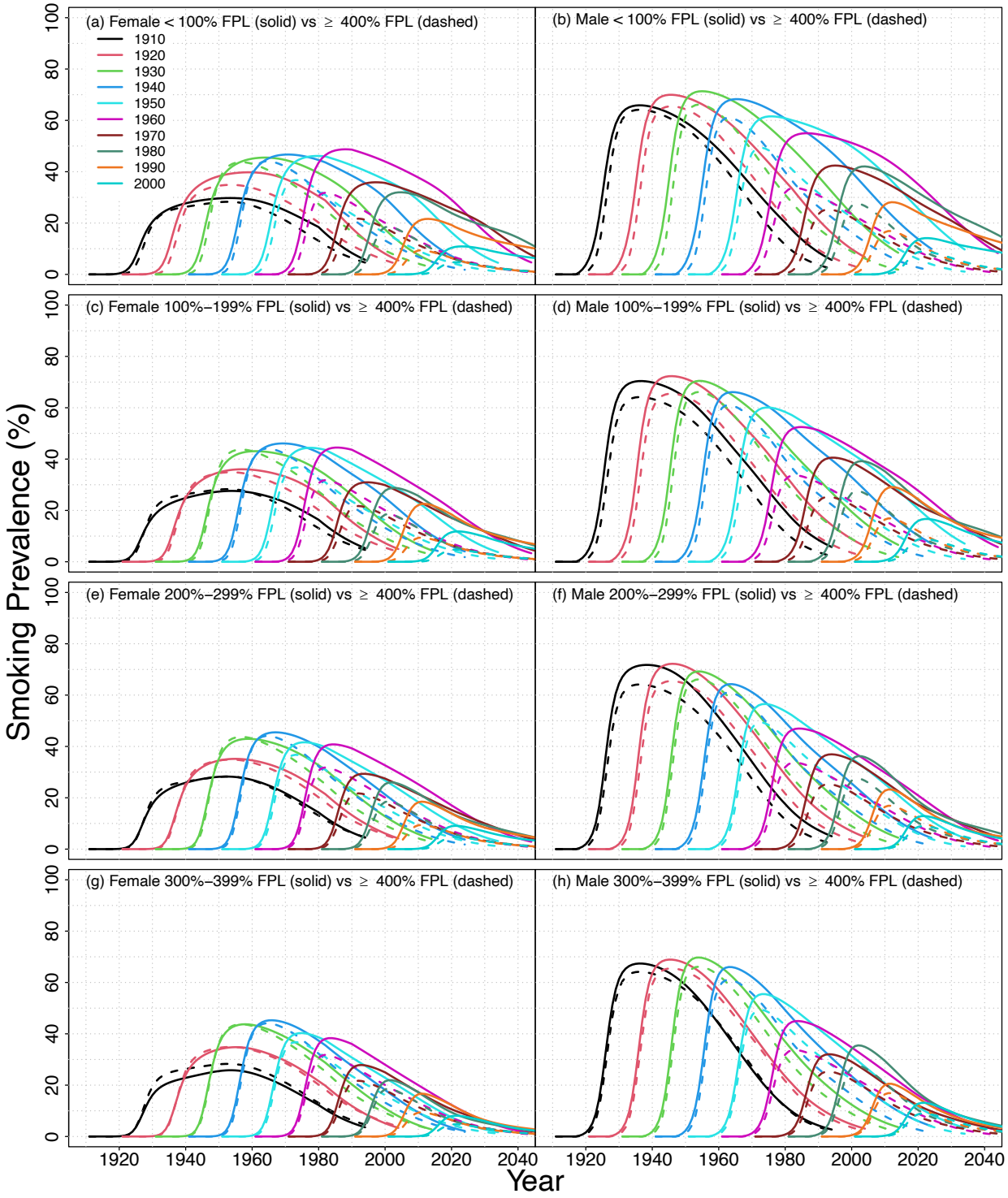

Figure S4. Age-specific smoking prevalence (percentage) for selected birth cohorts by family income-to-poverty ratio and gender (females – left panels, males – right panels). Solid lines correspond to Below poverty (<100% FPL, panels a & b), Near poverty (100%-199% FPL, panels c & d), 200%-299% FPL (panels e & f), and 300%-399% FPL (panels g and h). Dashed lines represent the smoking prevalence for ≥400% FPL which are shown as reference in all panels. FPL=Federal Poverty Level. An interactive version of this figure's data can be found at: https://sph-umich.shinyapps.io/shgdisplayappincome/

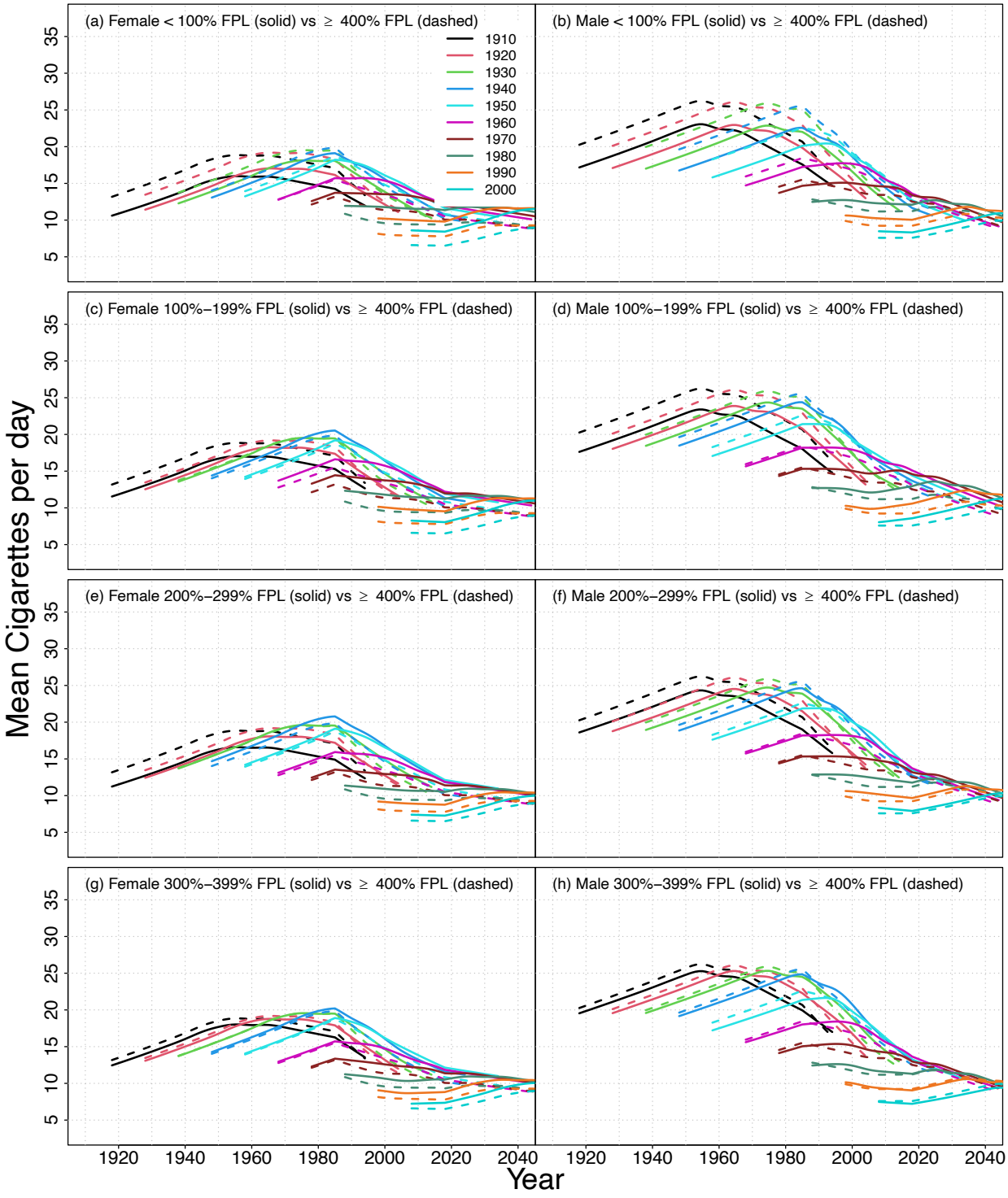

Figure S5. Age-specific mean cigarettes per day among current smokers for selected birth cohorts by family income-to-poverty ratio and gender (females – left panels, males – right panels). Solid lines correspond to Below poverty (<100% FPL, panels a & b), Near poverty (100%-199% FPL, panels c & d), 200%-299% FPL (panels e & f), and 300%-399% FPL (panels g and h). Dashed lines represent the mean CPD for ≥400% FPL which are shown as reference in all panels. FPL=Federal Poverty Level. An interactive version of this figure's data can be found at: https://sph-umich.shinyapps.io/shgdisplayappincome/

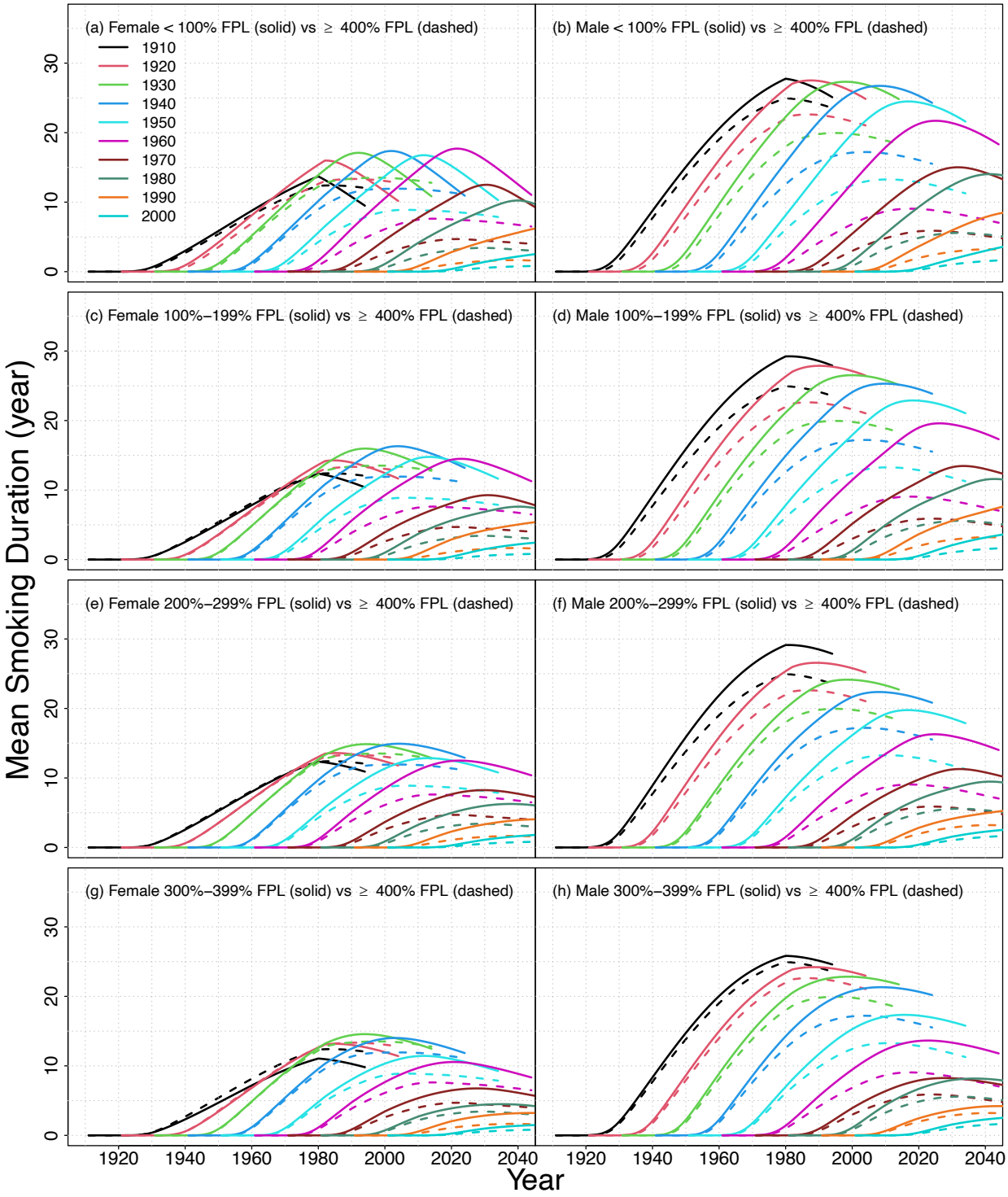

Figure S6. Age-specific mean smoking duration among smokers for selected birth cohorts by family income-to-poverty ratio and gender (females – left panels, males – right panels). Solid lines correspond to Below poverty (<100% FPL, panels a & b), Near poverty (100%-199% FPL, panels c & d), 200%-299% FPL (panels e & f), and 300%-399% FPL (panels g and h). Dashed lines represent the mean smoking duration for ≥400% FPL which are shown as reference in all panels. FPL=Federal Poverty Level. An interactive version of this figure's data can be found at: https://sph-umich.shinyapps.io/shgdisplayappincome/

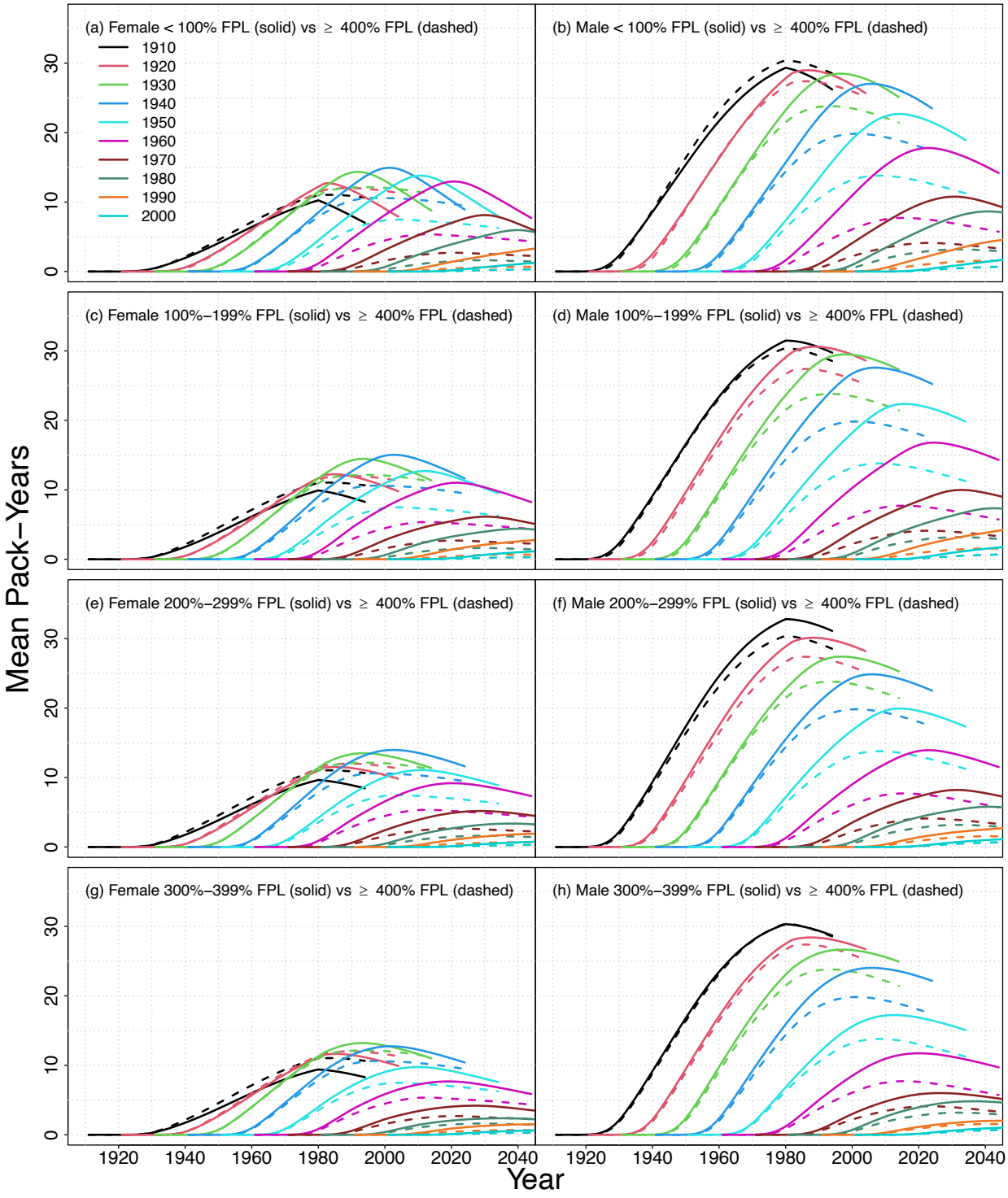

Figure S7. Age-specific mean pack-years among smokers for selected birth cohorts by family income-to-poverty ratio and gender (females – left panels, males – right panels). Solid lines correspond to Below poverty (<100% FPL, panels a & b), Near poverty (100%-199% FPL, panels c & d), 200%-299% FPL (panels e & f), and 300%-399% FPL (panels g and h). Dashed lines represent the mean pack-years for ≥400% FPL which are shown as reference in all panels. FPL=Federal Poverty Level. An interactive version of this figure's data can be found at: https://sph-umich.shinyapps.io/shgdisplayappincome/

S1 Text. Imputation for continuous family income in NHIS 1983-1996

The National Health Interview Survey (NHIS) survey provide information on detailed categories of family income and family income-to-poverty ratio, using the federal poverty level (FPL) from years 1997 and onwards.[1] The NHIS questionnaire was substantially redesigned in 2019, and it was discouraged to pool the data from 2019 and onwards with earlier years for any analysis since it would be difficult to detect whether any differences between surveys before and after 2019 are due to true change or changes in measurement and survey design.[2] Therefore, we used the NHIS imputed family income-to-poverty ratio data up to 2018, which was grouped in 14 categories: <50%; 50%-74%; 75%-99%; 100%-124%; 125%-149%; 150%-174%; 175%-199%; 200%-249%; 250%-299%; 300%-349%; 350%-399%; 400%-449%; 450%-499%; ≥500% of FPL.

In the 1983-1996 NHIS surveys, only a binary family income-to-poverty ratio data was available: at or above poverty threshold vs. below poverty threshold. Family income data were available with 27 categories in dollars: < 1,000; 19 categories from 1,000 to 19,999 with 1,000 increment (1,000-1,999, …, 19,000-19,999); 20,000-24,999; 25,000-29,999; 30,000-34,999; 35,000-39,999; 40,000-44,999; 45,000-49,999; ≥ 50,000. The missing rates for this categorical family income variable vary depending on the survey year: 12.1% (1983), 12.6% (1984), 12.3% (1985), 13.4% (1986), 13.5% (1987), 15.1% (1988), 16.4% (1989), 17.7% (1990), 18.9% (1991), 19.9% (1992), 16.9% (1993), 18.1% (1994), 17.0% (1995), 17.9% (1996). To compute the family income-to-poverty ratio, we imputed continuous family income from years 1983 to 1996.

NHIS has consistently reported continuous family income from 1997 to 2018, and imputed missing continuous family income data using a sequential regression multivariate imputation (SRMI) method. Before imputing missing data for continuous family income for years 1983-1996 in NHIS, we first validated the SRMI method by replicating the imputed family income for years 1997-2018, then applied this method to the earlier years (1983-1996). Additionally, using dataset with complete information for family income in 1983-1996, we randomly selected people with similar missing rates in NHIS data (ranged 12%-18% depending on year) and set their family income as missing. Then we imputed the missing family income using IVEware[3] software based on variables in Tables S1 and S2, and confirmed that the imputed family income generally matched to their actual reported family income. These validations guarantees that our imputed family income dataset is appropriate to estimate smoking patterns by income using the age-period-cohort models.

Imputation for continuous family income

The imputation of continuous family income from 1983-1989 was done in three steps. 1) Impute missing categorical family income data using the SRMI method. This method was used for the imputation of continuous family income from 1997 onwards by the NHIS.[3,4] 2) Fit either a continuous Weibull or Gamma distribution function to the non top-coded categorical family income data for each survey year and select the distribution that fits best the data in each case. 3) For each individual in each survey from 1983-1989, impute continuous family income by sampling from the corresponding best fitted distribution to categorical family income data from step 2), given the individual's categorical income. Table S1 and S2 present person-level covariates and variables at family level included in the imputation for missing categorical family income in step 1. These variables are the subset of the variables used in the imputation for missing continuous family income data from years 1997 by the NHIS,[4] which were available in the NHIS public data for years 1983-1989.

For step 1, we imputed missing categorical income as follows. We first impute missing values for any person-level covariates for adults using the Module "IMPUTE" in the IVEware software,[3] which was developed based on the SRMI method. A proper distribution should be assigned depending on the variable type; continuous, binary, categorical with more than two categories, count, semi-continuous. Then create covariates at family level by aggregating person-level covariates. The last step was to impute the missing values of family income categories and any missing values of family-level covariates due to missing person-level covariates for children, using the Module "IMPUTE" in the IVEware. This imputation approach was adopted from the NHIS imputation of continuous family income for 1997-2018.[4]

For step 2, using the "fitdist" function in R package "fitdistrplus", we fit either Weibull or Gamma distribution to the imputed family income categorical data from step 1. We chose a distribution which provided a smaller Akaike information criterion in each survey.

For the final step 3 (imputation of continuous family income), we imputed a continuous family income by sampling from the corresponding best fitted distribution to categorical family income data in step 2, given the individual's income category if it was below $50,000. For the top-coded family income category, i.e., ≥ $50,000, we sampled a continuous family income using the Pareto distribution with a location at $50,000 and a shape value (α) of $\log_4 5 \approx 1.16$.[5,6]

The NHIS provided imputed categorical family income data for 1990-1996, therefore, we skipped the step 1 when imputing continuous family income data for these years.

Computation of family income-to-poverty ratio

Given individuals' family size and structure (i.e., number of adults and related children under 18 years of age), we obtained the corresponding federal poverty level (FPL) using the annual U.S. Federal Poverty Guidelines.[7] With individuals' FPL and imputed continuous family income, we calculated the family income-to-poverty ratio for years 1983-1996. To reduce uncertainty that may arise from imputation, we did multiple imputation and obtained five sets of continuous family income data for 1983-1989, which is consistent to the number of sets for imputed family income-to-poverty ratio data reported in the NHIS for 1997-2018. For 1990-1996, the NHIS provided one set of imputed categorical family income data, and we imputed the continuous family income by using the steps 2 & 3 described above. For the age-period-cohort analyses by the level of family income, for each individual in years 1983-1989, we took an average over five family income-to-poverty ratios. Similarly, for each individual in years 1997-2018, we converted the NHIS imputed family income-to-poverty ratio categories into continuous income-to-poverty ratio by taking mid-point of each category, then took average over these five sets of income-to-poverty ratio. Then we categorized this continuous income-to-poverty ratio into five categories: below poverty (<100% FPL), near poverty (100%-199% FPL), 200%-299% FPL, 300%-399% FPL, and ≥400% FPL.

S2 Text. Methodology Overview for the Age-Period-Cohort Model

In this supplementary material we provide mathematical details on how the initiation and cessation probabilities, smoking prevalence, and smoking intensity (cigarettes per day) by family income were estimated using age-period-cohort (APC) models. Various age-period-cohort approaches have been proposed to estimate period and cohort trends. [8–16] Here we apply the approach by Holford et al.[15,16] to estimate age-specific smoking initiation, cessation, prevalence and intensity by cohort for different US income groups. These are estimable functions of the parameters,[17] thus, unaffected by the identifiability problem associated with APC models. We fitted consistent age-period-cohort models with consistent modeling parameters (spline knots etc.) for each income group to obtain income-specific smoking parameters. The details of each model fitted are provided below.

A compartment model that characterizes a typical smoking history is shown in Figure S8, in which a subject begins to smoke at some point after which they may quit. While this oversimplifies 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, and the individual is reclassified as current smoker at their reported cessation age.

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.,

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

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

$$P_E(a, c) = 1 - \prod_{i=1}^{a}[1 - p(i, c)] = 1 - [1 - P_E(a-1, c)][1 - p(a, c)] \qquad (1)$$

where $P_E(0, c) = 0$, 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, of course, mortality is affected by smoking so that the observed proportion of the population who have ever smoked at a particular age is given by $P_E^*(a, c) \leq P_E(a, c)$. 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 $C_p \geq 1$ 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 that the impact intensified with age. For example, cohorts born before 1953 would only have survey data for ages over 30 when one might expect differential mortality to begin to introduce substantial bias in the unadjusted estimate, $\hat{p}^*(a, c)$. In recent cohorts, almost all smoking initiation occurred before age 30, but for those born early in the twentieth century it was not so uncommon for initiation to occur later in life, especially in females. Later smoking initiation would also tend to postpone the effect of differential mortality in the cohort. We assumed that the differential mortality resulting from cigarette smoking occurred at ages, $a \geq a_0$, and $P_E^*(a, c) = P_E(a, c)$ 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 1983 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

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

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

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

For simplicity, we assumed that this quantity does not depend on the age an individual 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

$$P_F(a,c) = P_E(a,c) - P_C(a,c) = P_E(a,c)[1 - Q(a,c)]$$

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.,

$$P_C(a,c) + P_F(a,c) + P_N(a,c) = 1$$

Estimation of smoking parameters

Data were only obtained for a restricted range of ages, $a \in [\underset{.}{a}, \dot{a}]$, and periods, $t \in [\underset{.}{t}, \dot{t}]$ so that the earliest cohort would be $\underset{.}{c} = \underset{.}{t} - \dot{a}$ and the latest $\dot{c} = \dot{t} - \underset{.}{a}$. Available data, for a given cohort c, would cover an age range that would vary by cohort, i.e., $a \in [t - c, \dot{t} - c]$. 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 1908 to 2000 have been considered in the model. The earliest birth cohort (1908) is represented in the 1983 survey by subjects 75 and older. Because survey participants must be at least 18, the latest cohort was born in 2000 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.

Cross-sectional estimates of ever smokers by family income

For years with smoking status and family income surveyed, i.e., 1983, 1985, 1987, 1988, 1990-1995, and 1997-2018, 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 Y_i 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, $P_E(a,c)$. We assume an additive logistic model for Y_i, so that

$$\text{logit}\{P_E(a,c)\} = \beta_0 + \beta_a(a) + \beta_c(c)$$

where β_0 is an intercept and $\beta_\cdot(\cdot)$ is a function given by a constrained natural spline.[18] The model was fitted using PROC SURVEYLOGISTIC in SAS® with knots specified as

Age: 40, 50, 60, 70

Cohort: 1930, 1940, 1950, 1960, 1970, 1980, 1990

Age effects were estimated between 30 and 99. We fixed the cohort effect in earlier birth cohorts to be the same as those born before 1908, the earliest birth cohort that would provide data to a survey regarding smoking history after age 75 in 1983.

Smoking initiation probability

Unadjusted estimates of age-specific smoking initiation probabilities for a given cohort, $\hat{p}^*(a,c)$, were directly derived from the NHIS data. Information on age at start and family income was collected at survey years: 1987, 1988, 1992, 1995, and 1997-2018. 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.,

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

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

To determine the crude initiation probability estimates, an age-period-cohort model with period effect constraint to be 0 with 0 slope was fitted to the tabulated data given number of subjects who start smoking and are at risk of starting at a given age,

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

where β_0 is an intercept and $\beta_\cdot(\cdot)$ are given by constrained natural splines. 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.[18] Knots were specified as:

age: 10, 13, 16, 19, 22, 50, 60

period: 1940, 1950, 1960, 1970, 1980, 1990, 2000

cohort: 1920, 1930, 1940, 1950, 1960, 1970, 1980, 1990

Age for the target used to determine the correction factor was age in 1983 (year of the first NHIS survey when we imputed continuous family income data) or 30, whichever was older, $a^* = max\{1983 - c, 30\}$. The target value for the cumulative probability of being a smoker was the estimate derived in the analysis of the prevalence curve, $\widehat{\Pi}(a^*, c)$. Period and cohort effects before 1908 were assumed the same at the level of 1908, i.e., $\beta_t(t) = \beta_t(1908)$ for t < 1908, and $\beta_c(c) = \beta_c(1908)$ for c < 1908.

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 in years 1983, 1985, 1990, 1992, 1994, 1995, and 1997–2018, 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 with cohort effects fixed to be 0 with 0 slope was fitted in which

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

where β_0 is an intercept and $\beta_\cdot(\cdot)$ are given by 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.[18] Knots were specified as follows:

age: 30, 40, 50, 60

period: 1940, 1950, 1960, 1970, 1980, 1990, 2000

cohort: 1930, 1940, 1950, 1960, 1970, 1980, 1990

Estimates of the yearly cessation probability for age a and cohort c were the fitted values for ages 15-99, $\hat{q}(a,c)$. The conditional cessation probabilities were used to generate the cumulative probabilities of not quitting, $\hat{Q}(a,c)$, using equation (2). Period effects before 1923 were assumed the same as those in 1923, which is represented in the 1983 survey by subjects 60 and older. Cohort effects before 1908 were assumed the same as those in 1908.

Cigarettes smoked per day

Reports of the number of cigarettes smoked per day (CPD) showed an extremely high degree of digit preference, especially concentrated at half or whole 20 cigarette packs. Therefore, consistent with previous work,[15,16] we analyzed dose 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) were as following: CPD≤5 (3); 5< CPD≤15 (10); 15<CPD≤25 (20); 25<CPD≤35 (30); 35<CPD≤45 (40); and 45<CPD (60). Cigarettes per day information by family income are available in survey years: 1983, 1985, 1988, 1991-1995, and 1997-2018.

Constrained age effects based on estimates for the whole US population (all income groups)

Since the CPD estimates by family income may not be stable due to limited sample size in each CPD and income group, we fixed and used the same CPD age effects for all income groups, which were previously estimated using data from all respondents in the NHIS (all-income analysis). To obtain the all-income age effects, a multinomial logistic model was fitted to the NHIS 1965-2018 data using PROC SURVEYLOGISTIC with reported sample adult weights in SAS® with age, period and cohort represented by additive nonparametric factors function of time using constrained natural splines. Yearly estimates of the linear multinomial logistic age-period-cohort model with period effects fixed to be 0 with 0 slope was fitted in which

$$\log\left\{\frac{P(CPD_{a,c}=k)}{P(CPD_{a,c}=1)}\right\} = \beta_0 + \beta_a(a) + \beta_t(t) + \beta_c(c)$$

where k=2, 3, …, 6, representing the CPD category specified above with 6 indicating the most intensive CPD category, β_0 is an intercept and $\beta.(\cdot)$ are given by constrained natural splines. Knots were specified as:

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

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

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

Cigarettes per day by family income

We fitted an age-period-cohort model with the all-income age effects as an offset. The knots were specified below:

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

Period knots: 1990, 2000, 2005, 2010

Cohort knots: 1930, 1940, 1950, 1960, 1970

Period effects before 1985 were fixed as the estimates in 1985 and after 2018 fixed as the ones in 2018. Similarly, cohort effects before 1920 were fixed as the estimates in 1920 and after 1980 fixed as the ones in 1980.

Mean cumulative years smoked and mean pack-years

We estimated the mean cumulative years smoke, i.e., mean smoking duration (MSD) and the mean pack-years (MPKY) by family income using the formulas presented below. We used the previously estimated initiation and cessation probabilities, current/former smoker prevalence and cigarettes per day by age, birth cohort and gender to calculate the MSD and MPKY.

For current smokers aged *a* in birth cohort *c*, with initiation age *t* ($0 \le t \le a$), we first determine the distribution of *t* by finding the cumulative distribution function of t conditional on a and c: $F_I(t|a,c)$. $F_I(0|a,c) = 0$ and

$$F_I(t|a,c) = 1 - [1 - F_I(t-1|a,c)][1 - p(t-1|a,c)], for\ 0 < t \le a$$

Thus, the probability distribution function of initiation age is

$$f_I(t|a,c) = \frac{F_I(t|a,c) - F_I(t-1|a,c)}{F_I(a|a,c)}$$

Smoking duration at age *a* for an individual who begins smoking at age *t* is *(a-t)*, assuming everything happens at the midpoint. So the mean duration of all current smokers at age *a* is

$$\mu_c(a|c) = \sum_{i=0}^{a} f_I\,(i|a,c)(a-i)$$

Similarly, if *x(t|c)* is the estimated mean cigarettes per day at age *t* for cohort *c*, the MPKY of exposure for current smokers at age *a* is

$$\theta_c(a|c) = \sum_{i=0}^{a} f_I\,(i|a,c) \sum_{k=i}^{a} x\,(k|c) * g(k)/20,\ where\ g(k) = \begin{cases} 0.5 & \text{if k=i or k=a} \\ 1 & \text{otherwise} \end{cases}$$

For former smokers aged *a* in birth cohort *c*, with initiation age at *t* and cessation age at t_q, the distribution of initiation age is the same as that of current smokers. For former smokers, we need to also define the cumulative distribution function of cessation given age a in birth cohort c, denoted as $F_Q(t_q|a,c,t)$,

$$F_Q(t_q|a,c,t) = 1 - [1 - F_Q(t_q - 1|a,c,t)][1 - q(t_q - 1|a,c,t)],\ for\ t \le t_q \le a$$

The probability density function of t_q is

$$f_Q(t_q|a,c,t) = \frac{F_Q(t_q|a,c,t) - F_Q(t_q - 1|a,c,t)}{F_Q(a|a,c,t)}$$

The MSD for former smokers aged *a* given birth cohort *c* is then calculated as

$$\mu_f(a|c) = \sum_{i=0}^{a}\sum_{j=i}^{a} f_I\,(i|a,c) f_Q(j|a,c,i)(j-i) = \sum_{i=0}^{a} f_I\,(i|a,c) \sum_{j=i}^{a} f_Q\,(j|a,c,i)(j-i)$$

The MPKY for former smokers aged *a* born in birth cohort *c* is calculated as

$$\theta_f(a|c) = \sum_{i=0}^{a}\sum_{j=i}^{a} f_I\,(i|a,c) f_Q(j|a,c,i) \sum_{k=i}^{j} x\,(k|c) * g(k)/20,\ where\ g(k) = \begin{cases} 0.5 & \text{if k=i or k=j} \\ 1 & \text{otherwise} \end{cases}$$

Finally, the overall MSD for all smokers aged *a* in birth cohort *c* is calculated as

$$\mu_a(a|c) = \mu_c(a|c) * P_C(a,c) + \mu_f(a|c) * P_F(a,c),$$

where $P_C(a,c)\ and\ P_F(a,c)$ are the prevalence of current and former smokers at age a in birth cohort c, respectively. We can derive the MPKY for all smokers at age a in birth cohort *c* similarly as

$$\theta_a(a|c) = \theta_c(a|c) * P_C(a,c) + \theta_f(a|c) * P_F(a,c)$$

Figure S8. Compartments considered in developing smoking history.

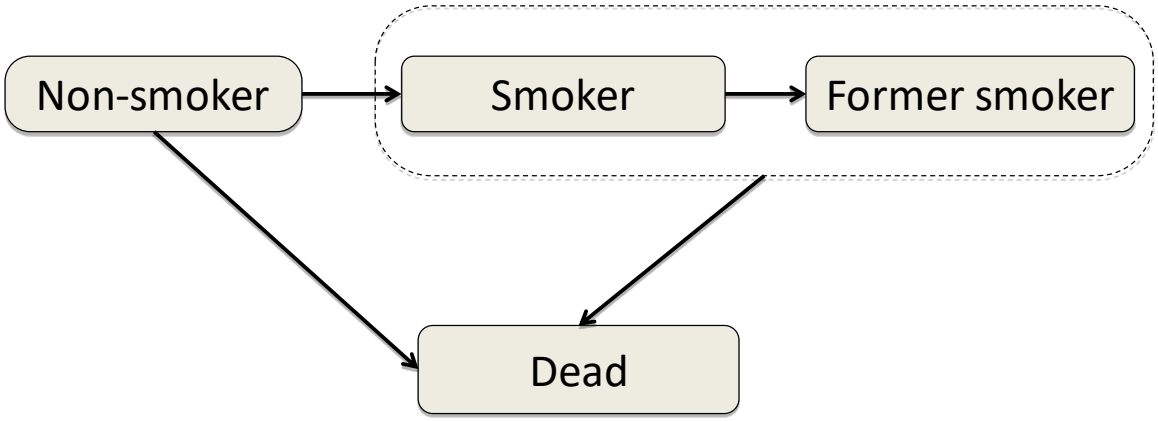

Table S1. Variables included in imputation of person-level covariates in NHIS for 1983-1989.

Variable Name	Description and Code Values (NHIS variable*)
SEX	Sex 1 = Male, 2 = Female
AGEGROUP	Age group (AGE) 0 = under 18, 1 = 18-24, 2 = 25-44, 3 = 45-64, 4 = 65+
ORIGIN	Ethnic origin (HISPETH) 1 = Hispanic, 2 = Non-Hispanic
RACEREC	Racial background (RACEA) 1 = White, 2 = Black, 3 = Other
MARRY	Current marital status (MARST) 1 = Married, 2 = Widowed/Divorced/Separated, 3 = Never married, 4 = Under 14 years old
FM_SIZER	Family size (FAMSIZE) 1 = 1 person family, 2 = 2 person family, 3 = 3 person family, 4 = 4+ person family
WTFA	Final person weight (PERWEIGHT)
STRATA	Stratum for variance estimation
PSU	Primary sampling unit for variance estimation
LIM_ACT	Any limitation of activity (LATOTAL) 1 = Age 18+, Limited, 2 = Age 18+, Not limited, 3 = Under 18 years old
PHSTAT	Health status (HEALTH) 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor
PHOSPYR	In a hospital overnight in the past 12 months (HOSPNGHTD) 1 = Yes, 2 = No
P10DVYR	Received health care from doctor 10+ times in the past 12 months (DV12) 1 = Yes, 2 = No
M_CARE	Medicare coverage (HIMCARE) 1 = Yes, 2 = No
MILITRY	Military health care coverage (HIMILANY) 1 = Yes, 2 = No
PRIVATW	Private insurance coverage through employment (HIPWORKR) 1 = Yes, 2 = No
PRIVATS	Private insurance coverage purchased directly (HIPBUYOWNR) 1 = Yes, 2 = No
EDUCR	Educational attainment (EDUCREC2) 1 = High school or less, 2 = HS graduate or equivalent, 3 = Some college 4 = College graduate, 5 = More than college
EMPSTAT	Major or usual activity in the past 12 months (MAJACT) 1 = Age 18+, worked for pay last year 2 = Age 18+, not worked for pay last year 3 = Under 18 years old
PSSI	Person received income from Supplement Security Income (GOTSSI) 1 = Yes, 2 = No

*Harmonized NHIS variables across years provided by the IPUMS Health Surveys.[1]

Table S2. Variables included in imputation at family-level for 1983-1989

Variable Name	Description and Code Values (NHIS variable*)
WTFA_FAM	Final family weight (HHWEIGHT)
STRATA	Stratum for variance estimation
PSU	Primary sampling unit for variance estimation
ADULT	Total number of adults in a family
CHILD	Total number of children in a family
M_ERNAGE	Average age of male earners in a family
F_ERNAGE	Average age of female earners in a family
FM_EARN	Total number of earners in a family
P_HISP	Proportion of Hispanics in a family
P_WHITE	Proportion of whites in a family
P_BLACK	Proportion of blacks in a family
FM_HOSP	Family having family members with PHOSPYR = 1 (In a hospital overnight in the past 12 months) 1 = at least one family member has 2 = none of the family members has
FM_DVYR	Family having family members with P10DVYR = 1 (Received health care from doctor 10+ times in the past 12 months) 1 = at least one family member has 2 = none of the family members has
FM_MCARE	Family having family members with M_CARE = 1 (Medicare coverage) 1 = at least one family member has 2 = none of the family members has
FM_MILIT	Family having family members with MILITRY = 1 (Military health care coverage) 1 = at least one family member has 2 = none of the family members has
FM_PRIVW	Family having family members with PRIVATW = 1 (Private insurance coverage through employment) 1 = at least one family member has 2 = none of the family members has
FM_PRIVS	Family having family members with PRIVATS = 1 (Private insurance coverage purchased directly) 1 = at least one family member has 2 = none of the family members has
FM_HLTH1	Family having family members with PHSTAT = 1 or 2 (Excellent or very good health) 1 = at least one family member has 2 = none of the family members has
FM_HLTH2	Family having family members with PHSTAT = 3 (Good health) 1 = at least one family member has 2 = none of the family members has
FM_HLTH3	Family having family members with PHSTAT = 4 or 5 (Fair or poor health) 1 = at least one family member has 2 = none of the family members has
FM_HIEDU	Highest education attainment of family members 1 = High school or less 2 = HS graduate or equivalent 3 = Some college 4 = College graduate

	5 = More than college 6 = All family members are under 18 years old
FM_SSI	Family having family members with PSSI = 1 (Person received income from Supplement Security Income) 1 = at least one family member has 2 = none of the family members has
FM_LIM_ACT	Family having family members with LIM_ACT = 1 (Any limitation of activity) 1 = at least one family member has 2 = none of the family members has
FAMINC_IMP	Total family income category (in dollars) 1 = 0-3,999 2-17 = 16 categories with 1,000 increment (4,000-4,999, …, 19,000-19,999) 18-23 = 6 categories with 5,000 increments (20,000 - 24,999, …, 45,000 - 49,999) 24 = ≥ 50,000
P_F_EARN	Proportion of female earners to the total family earners

*Harmonized NHIS variables across years provided by the IPUMS Health Surveys.[1]

References

1. IPUMS NHGIS | National Historical Geographic Information System. Accessed January 13, 2022. https://www.nhgis.org/
2. IPUMS NHIS. Accessed February 18, 2022. https://nhis.ipums.org/nhis/userNotes_2019_NHIS_Redesign.shtml
3. IVEware: Imputation and Variance Estimation Software | Survey Research Center. Accessed December 23, 2021. https://www.src.isr.umich.edu/software/iveware/
4. Center for Health Statistics - Division of Health Interview Statistics N. Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples. Published online 2018.
5. Pareto principle - Wikipedia. Accessed February 3, 2022. https://en.wikipedia.org/wiki/Pareto_principle
6. Oancea B, Andrei T, Pirjol D. Income inequality in Romania: The exponential-Pareto distribution. *Physica A*. 2017;469:486-498. doi:10.1016/j.physa.2016.11.094
7. Prior HHS Poverty Guidelines and Federal Register References | ASPE. Accessed January 6, 2022. https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines/prior-hhs-poverty-guidelines-federal-register-references
8. Land KC, Yang Y. *Age-Period-Cohort Analysis : New Models, Methods, and Empirical Applications*. CRC Press; 2013.
9. van Hook J, Quirós S, Dondero M, Altman CE. Healthy Eating among Mexican Immigrants: Migration in Childhood and Time in the United States. *J Health Soc Behav*. 2018;59(3):391-410. doi:10.1177/0022146518788869
10. Fosse E, Winship C. Bounding Analyses of Age-Period-Cohort Effects. *Demography*. 2019;56(5):1975-2004. doi:10.1007/S13524-019-00801-6
11. Meza R, Hazelton WD, Colditz GA, Moolgavkar SH. Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage

carcinogenesis model. *Cancer Causes Control*. 2008;19(3):317-328. doi:10.1007/S10552-007-9094-5
12. Hazelton WD, Jeon J, Meza R, Moolgavkar SH. Chapter 8: The FHCRC lung cancer model. *Risk Anal*. 2012;32 Suppl 1(Suppl 1):S99-116. doi:10.1111/j.1539-6924.2011.01681.x
13. Jeon J, Luebeck EG, Moolgavkar SH. Age effects and temporal trends in adenocarcinoma of the esophagus and gastric cardia (United States). *Cancer Causes Control*. 2006;17(7):971-981. doi:10.1007/s10552-006-0037-3
14. Moolgavkar SH, Holford TR, Levy DT, et al. Impact of reduced tobacco smoking on lung cancer mortality in the United States during 1975-2000. *J Natl Cancer Inst*. 2012;104(7):541-548. doi:10.1093/jnci/djs136
15. Holford TR, Levy DT, McKay LA, et al. Patterns of birth cohort-specific smoking histories, 1965-2009. *Am J Prev Med*. 2014;46(2):e31-37. doi:10.1016/j.amepre.2013.10.022
16. Holford TR, Levy Phd DT, Meza R. Comparison of Smoking History Patterns Among African American and White Cohorts in the United States Born 1890 to 1990. *Nicotine Tob Res*. 2016;18 Suppl 1(Suppl 1):S16-29. doi:10.1093/ntr/ntv274
17. Holford TR. The Estimation of Age, Period and Cohort Effects for Vital Rates. *Biometrics*. 1983;39(2):311-324. doi:10.2307/2531004
18. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med*. 1989;8(5):551-561. doi:10.1002/SIM.4780080504