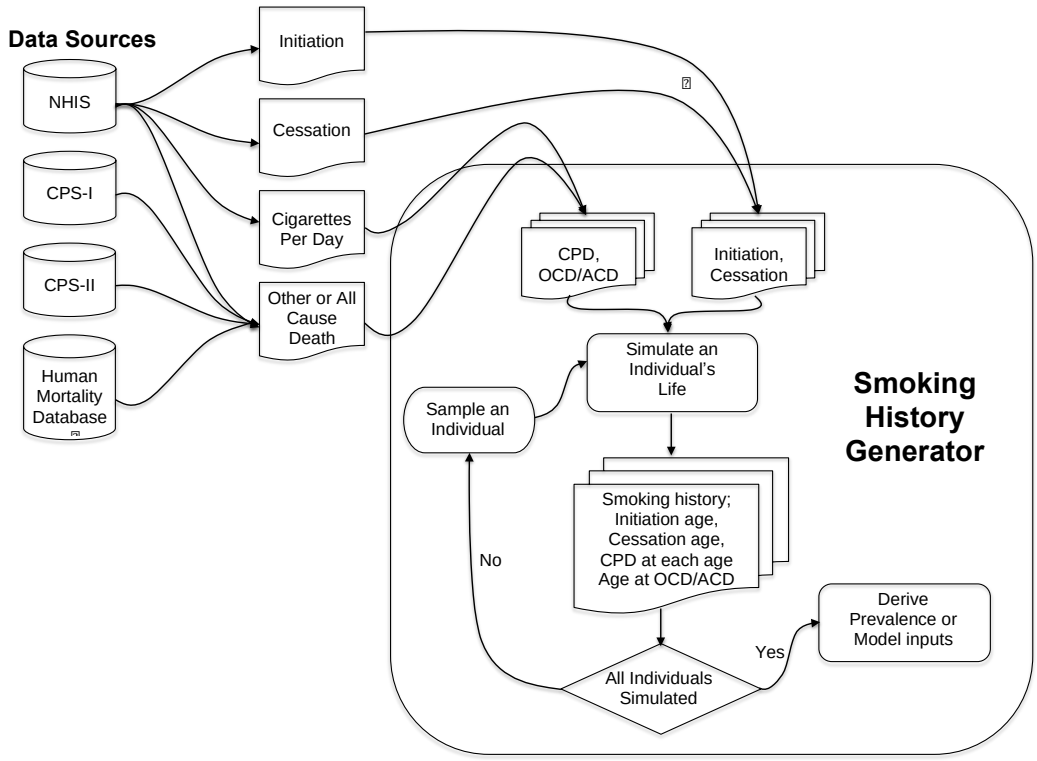
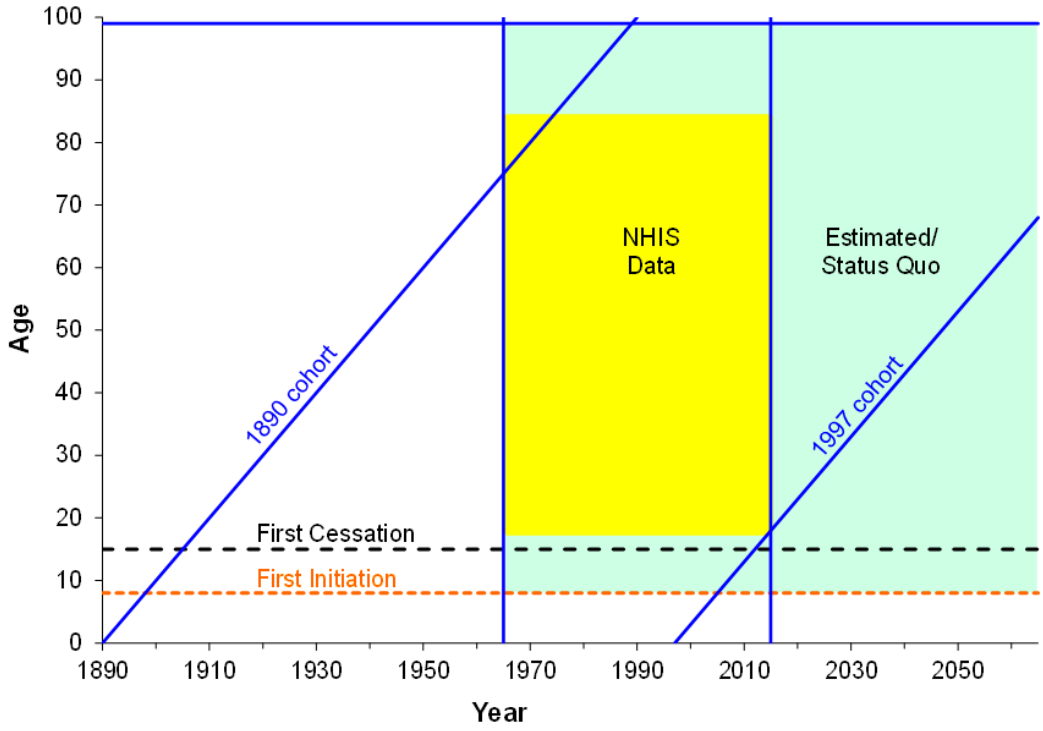
Supplementary Note to

"Smoking and Lung Cancer Mortality in the US from 2015-2065: a comparative modeling approach"

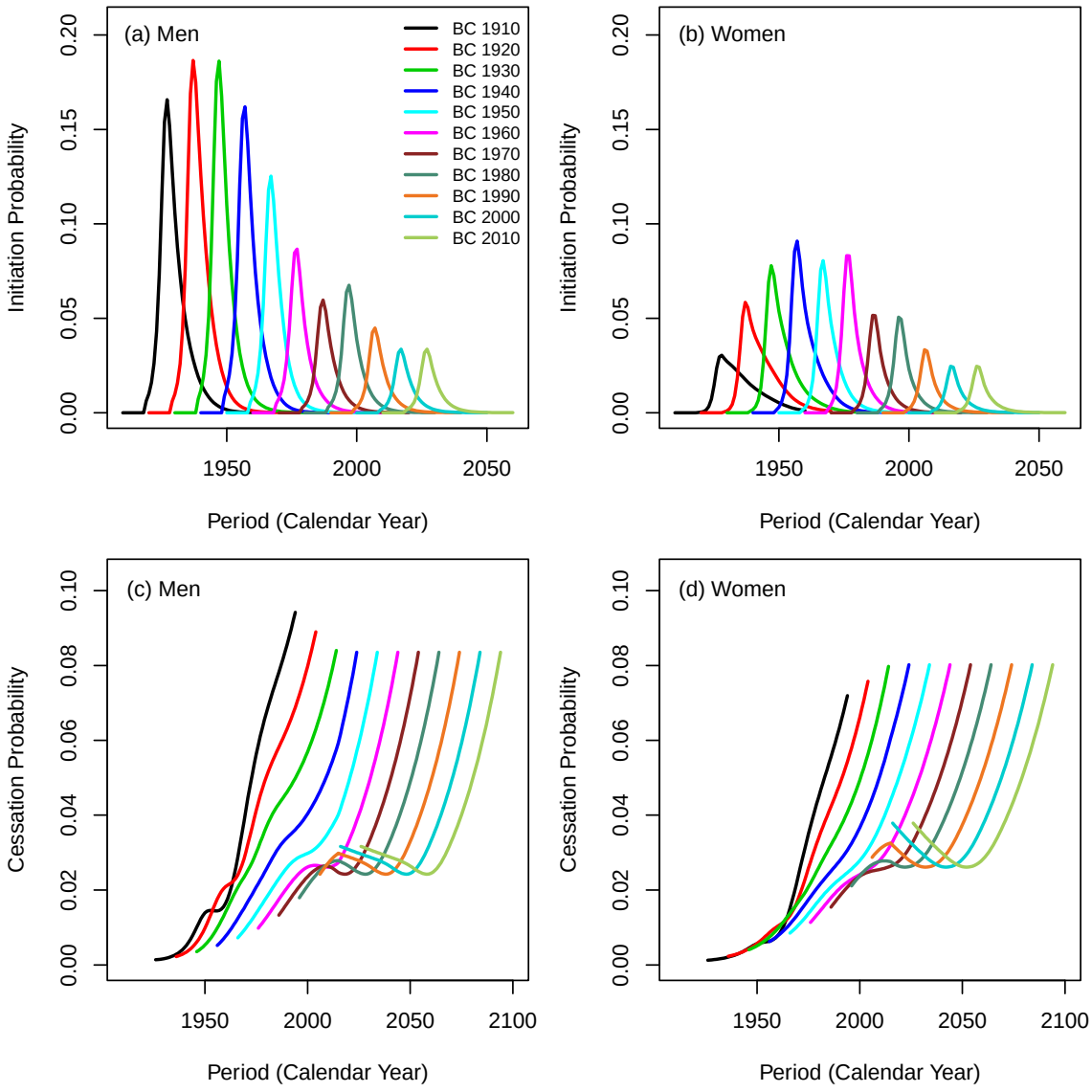
Jihyoun Jeon, Theodore R. Holford, David T. Levy, Eric J. Feuer, Pianpian Cao, Jamie Tam, Lauren Clarke, John Clarke, Chung Yin Kong, Rafael Meza



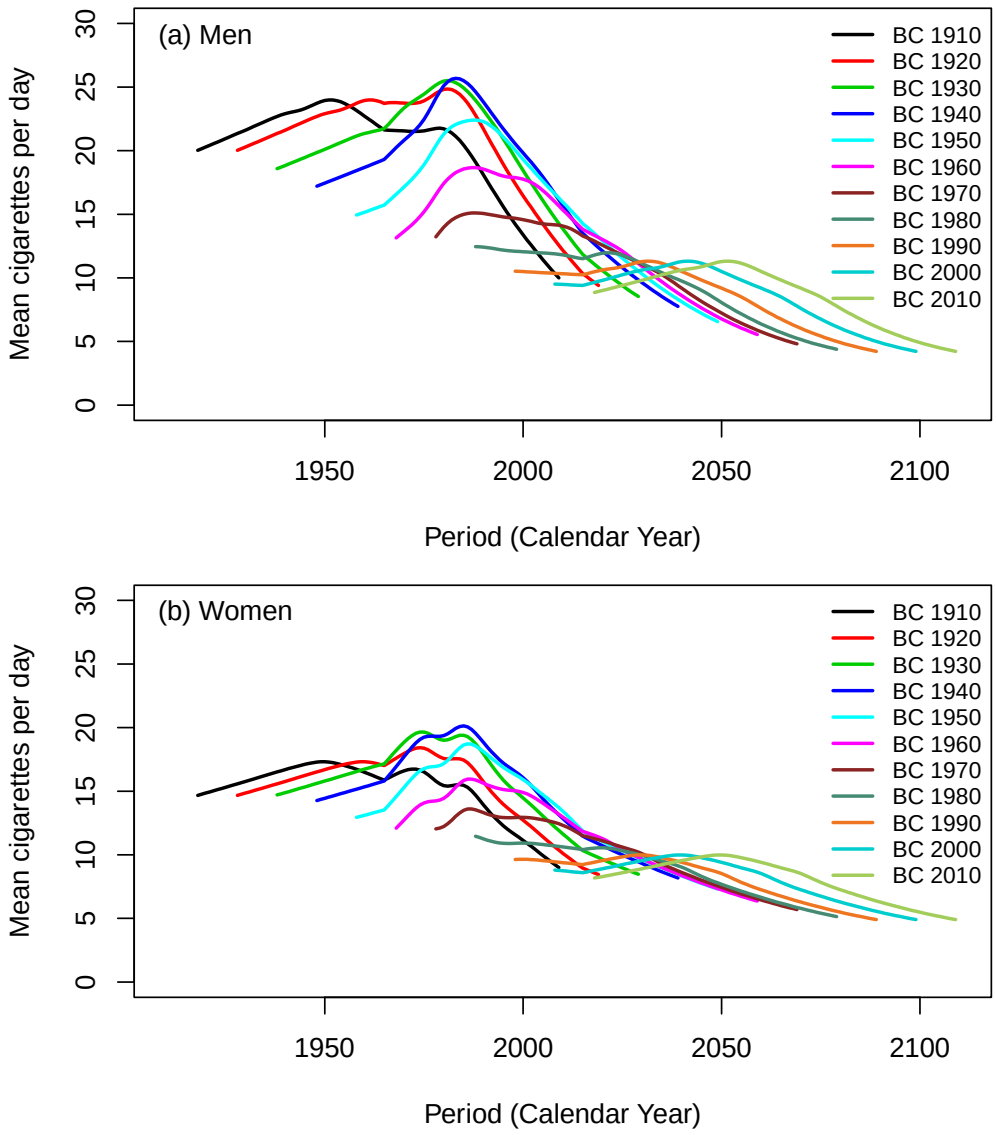
Appendix Figure 1. Flow-chart of the smoking history generator (SHG).



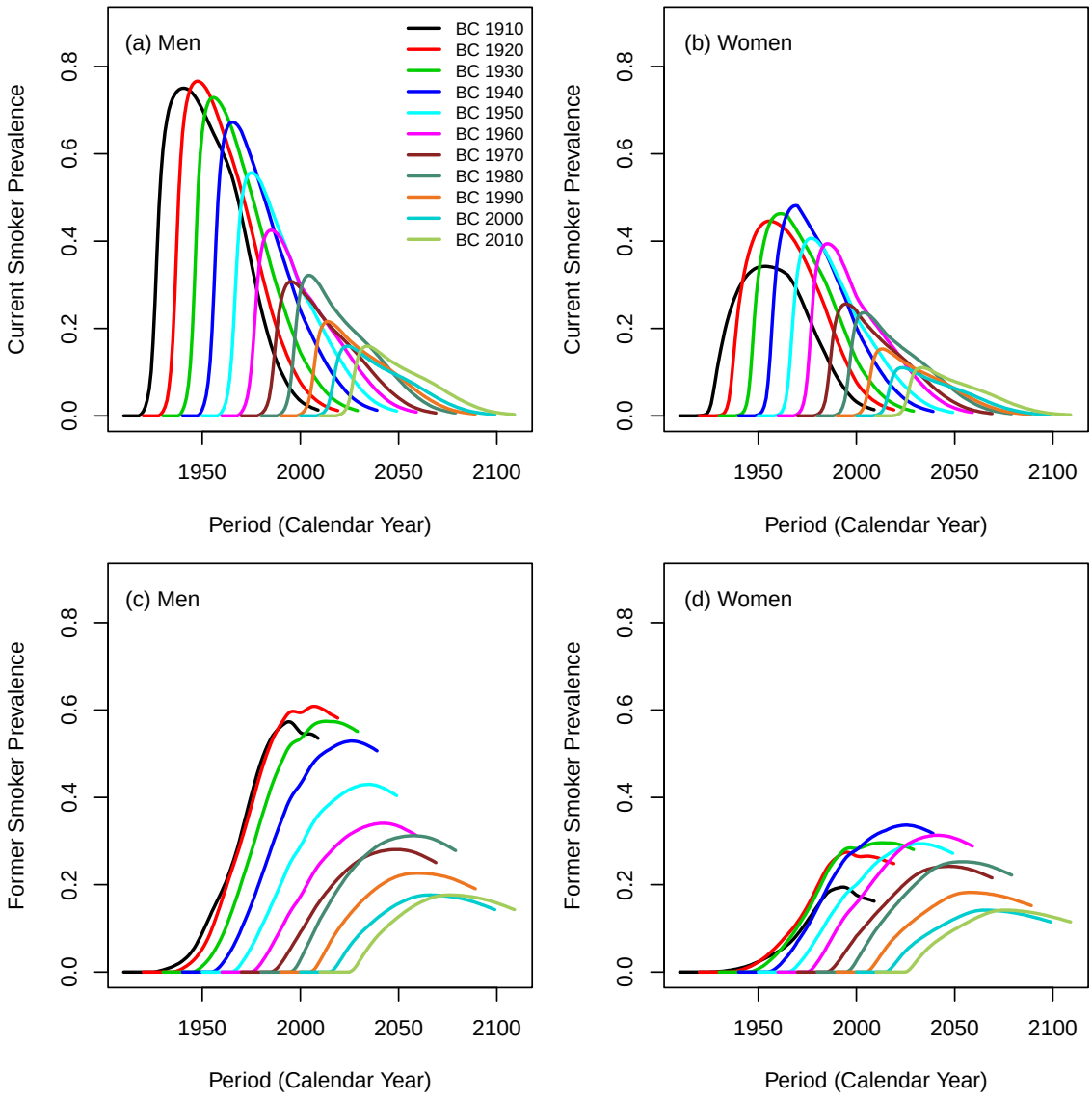
Appendix Figure 2. Smoking input data for the smoking history generator (SHG) under the status quo scenario. The area in yellow represents years and ages of smoking parameters covered by the underlying National Health Interview Survey (NHIS) data, and the area in green represents projected parameters using the fitted Age-Period-Cohort model.



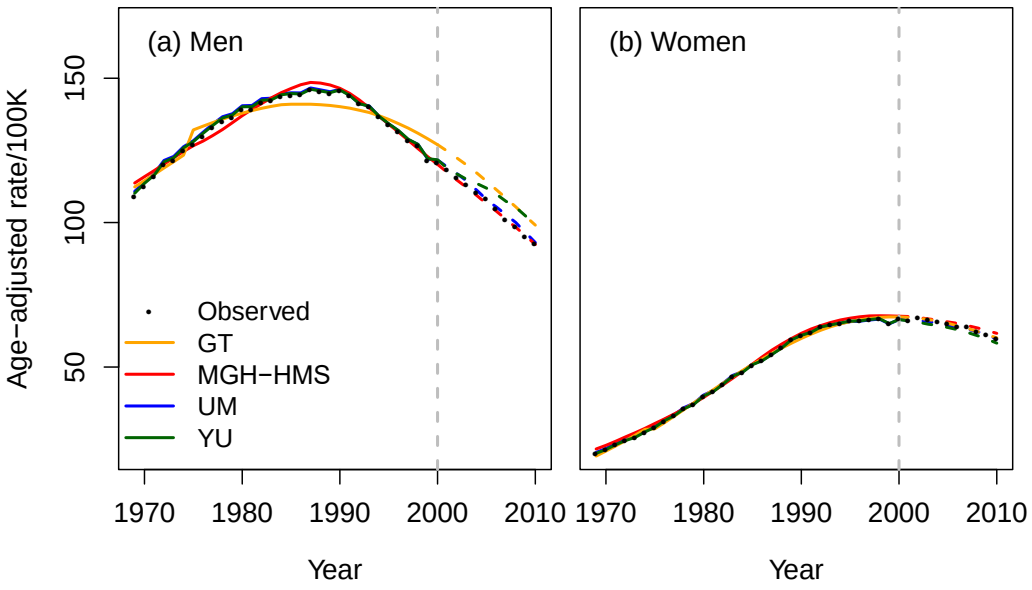
Appendix Figure 3. Annual smoking initiation and cessation probabilities for selected birth cohorts (BCs) under the status quo scenario by sex. An interactive version of the figure can be found at https://resources.cisnet.cancer.gov/projects/ - shg/sbc2/tool?figure=appendix_fig_3.



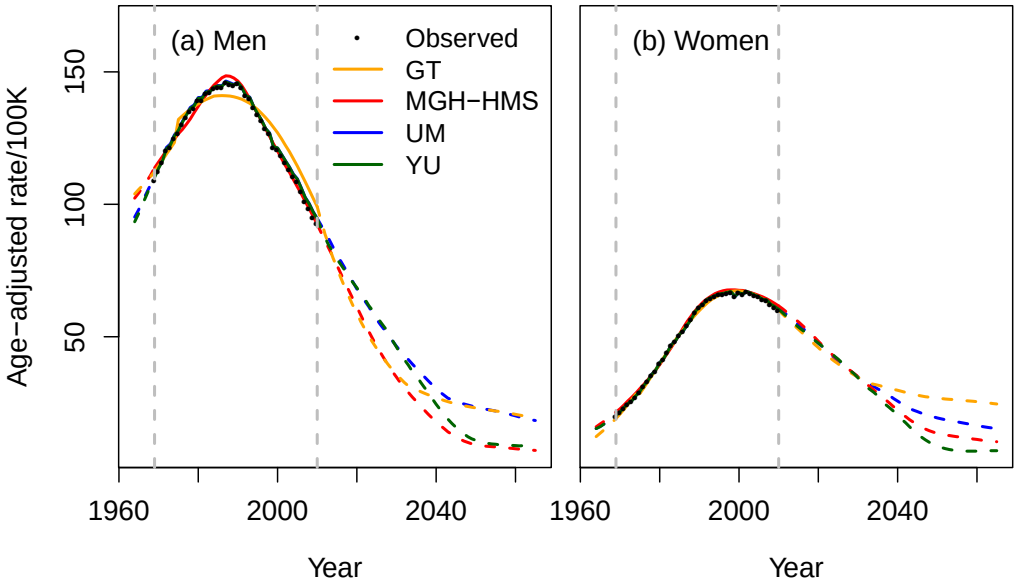
Appendix Figure 4. Annual mean cigarettes per day for selected birth cohorts (BCs) under the status quo scenario by sex. An interactive version of the figure can be found at https://resources.cisnet.cancer.gov/projects/ - shg/sbc2/tool?figure=appendix_fig_4.



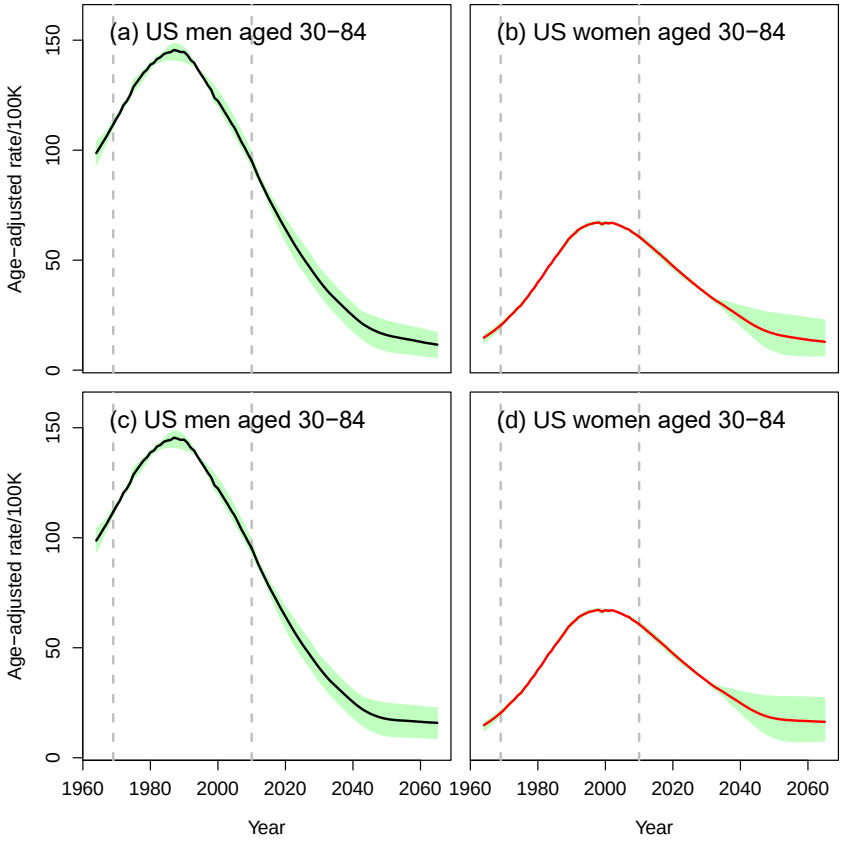
Appendix Figure 5. Prevalence of current and former smokers for selected birth cohorts (BCs) under the status quo scenario by sex. An interactive version of the figure can be found at https://resources.cisnet.cancer.gov/projects/ - shg/sbc2/tool?figure=appendix_fig_5.



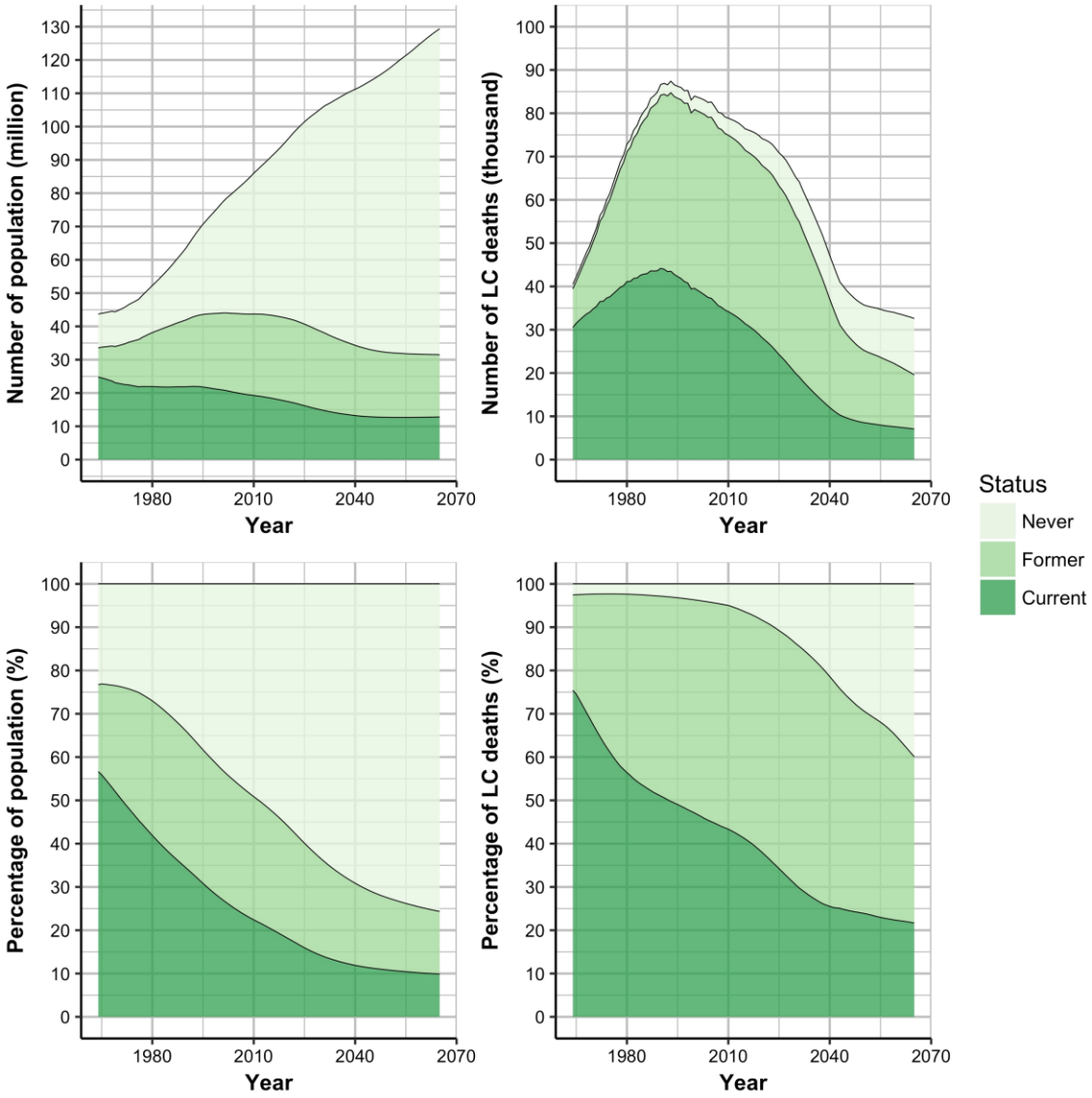
Appendix Figure 6. Age-adjusted lung cancer mortality rates per 100,000 over 1969-2010. Each model was calibrated to the observed US lung cancer mortality data from 1969-2000 (solid lines), and predicted lung cancer mortality rates for years 2001-2010 (dashed lines). The 2000 US standard population was used to calculate the age-adjusted rates. The black dots represent the observed US lung cancer mortality rates, and the lines prediction from four independent models: Georgetown University (GT), Massachusetts General Hospital and Harvard Medical School (MGH-HMS), University of Michigan (UM), Yale University (YU).



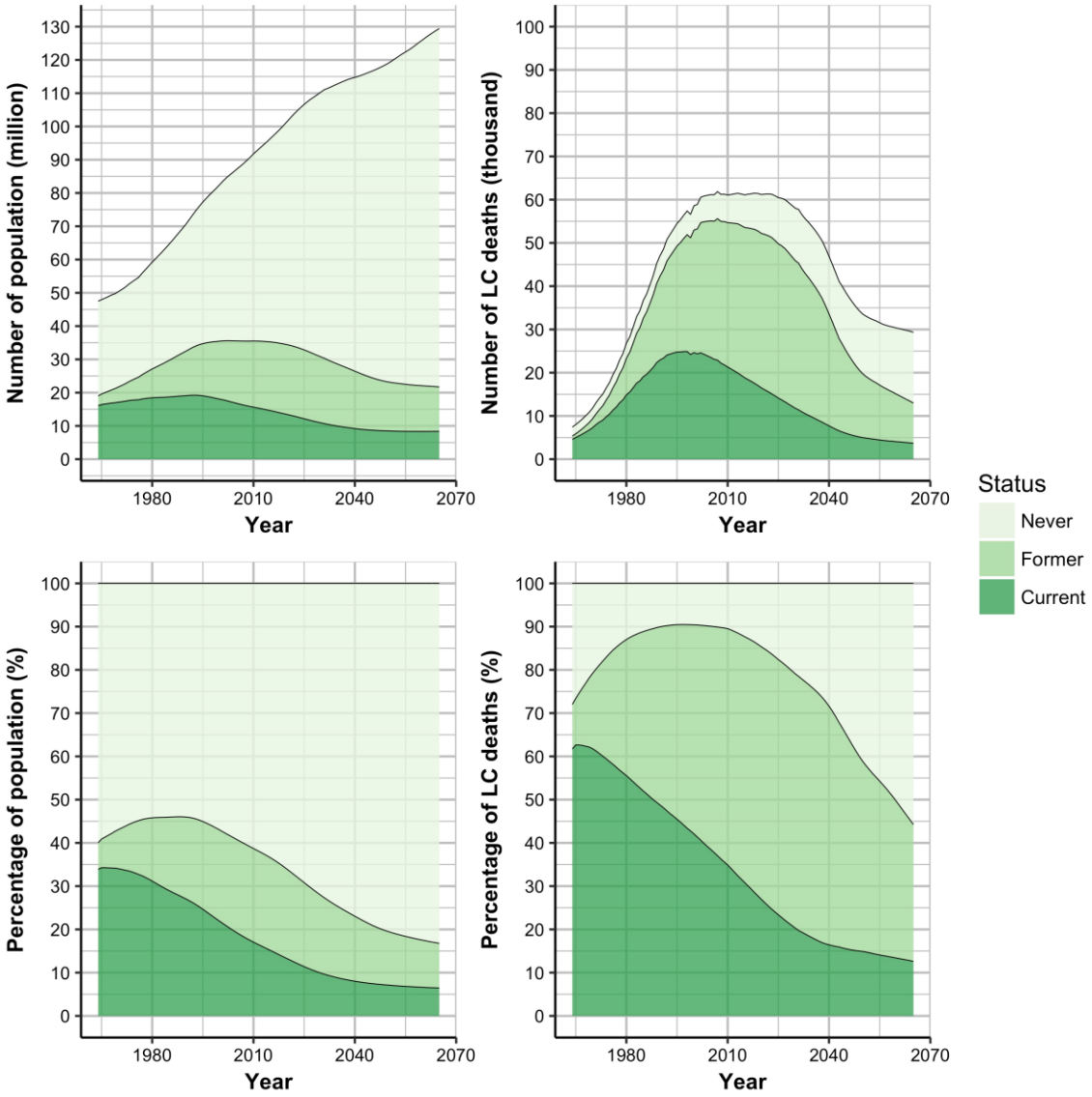
Appendix Figure 7. Age-adjusted lung cancer mortality rates per 100,000 for 1964-2065 under the status quo scenario. Each model (GT, MGH-HMS, UM and YU) was calibrated to the observed US lung cancer mortality data from 1969-2010 (solid lines), projected lung cancer mortality rates for 1964-1968 and 2011-2065 (dashed lines). The 2000 US population was used as the standard to calculate age-adjusted rates. The black dots represent the observed US lung cancer mortality rates, and the lines prediction from four independent models: Georgetown University (GT), Massachusetts General Hospital and Harvard Medical School (MGH-HMS), University of Michigan (UM), Yale University (YU).



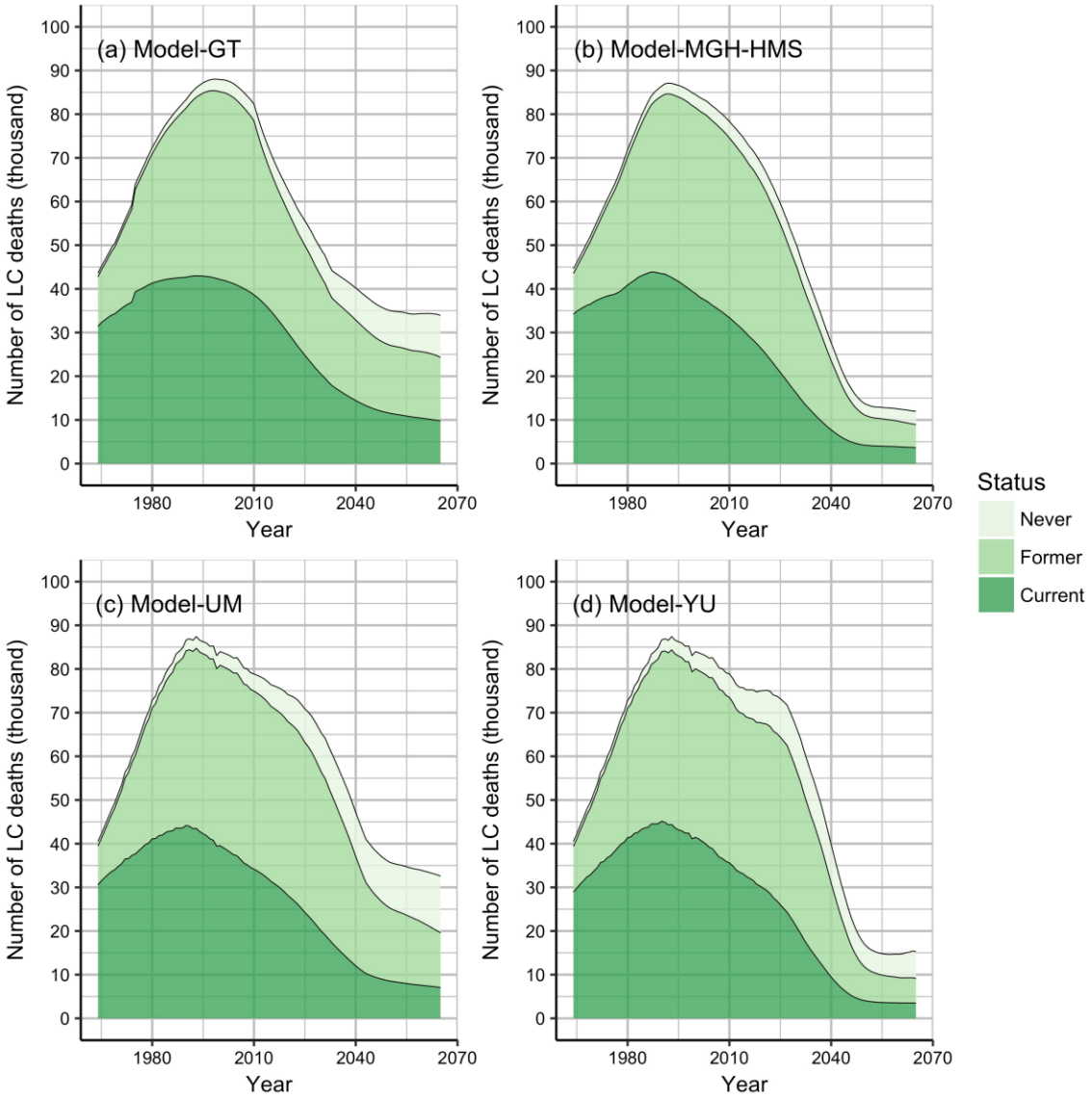
Appendix Figure 8. Age-adjusted lung cancer mortality rates per 100,000 for 1964-2065 under the optimistic, (a) & (b) and pessimistic, (c) & (d) scenarios. The line represents the mean age-adjusted lung cancer mortality rate across four CISNET-Lung models, and the shaded area shows the range of age-adjusted lung cancer mortality rates per 100,000 across four models. The 2000 US population was used as the standard to calculate age-adjusted rates.



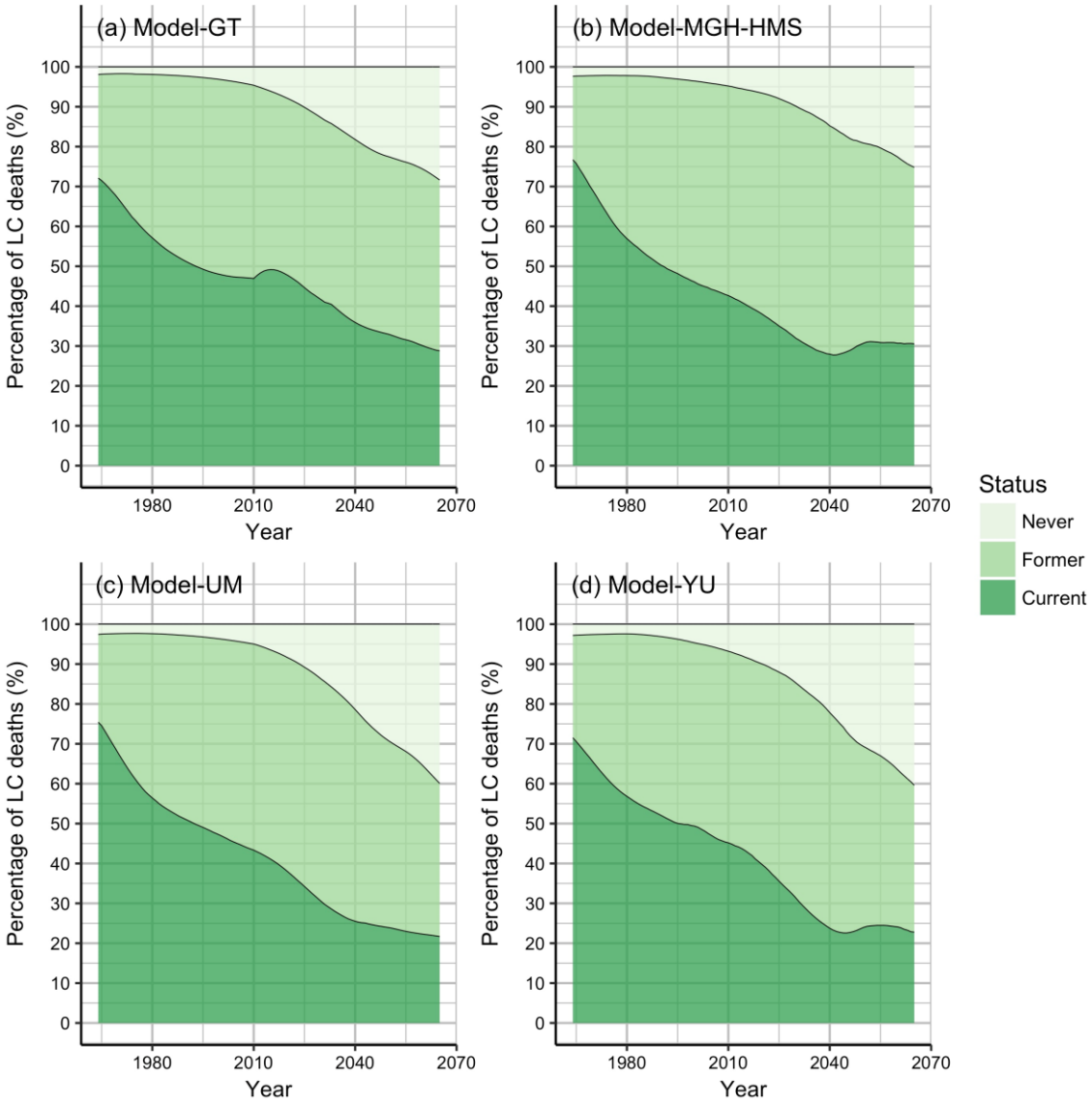
Appendix Figure 9. Population and lung cancer deaths from the University of Michigan (UM) model. Number and percentage of US men aged 30-84 by smoking status for 1964-2065 (left panels) and number and percentage of lung cancer deaths by smoking status for 1964-2065 (right panels), under the status quo scenario.



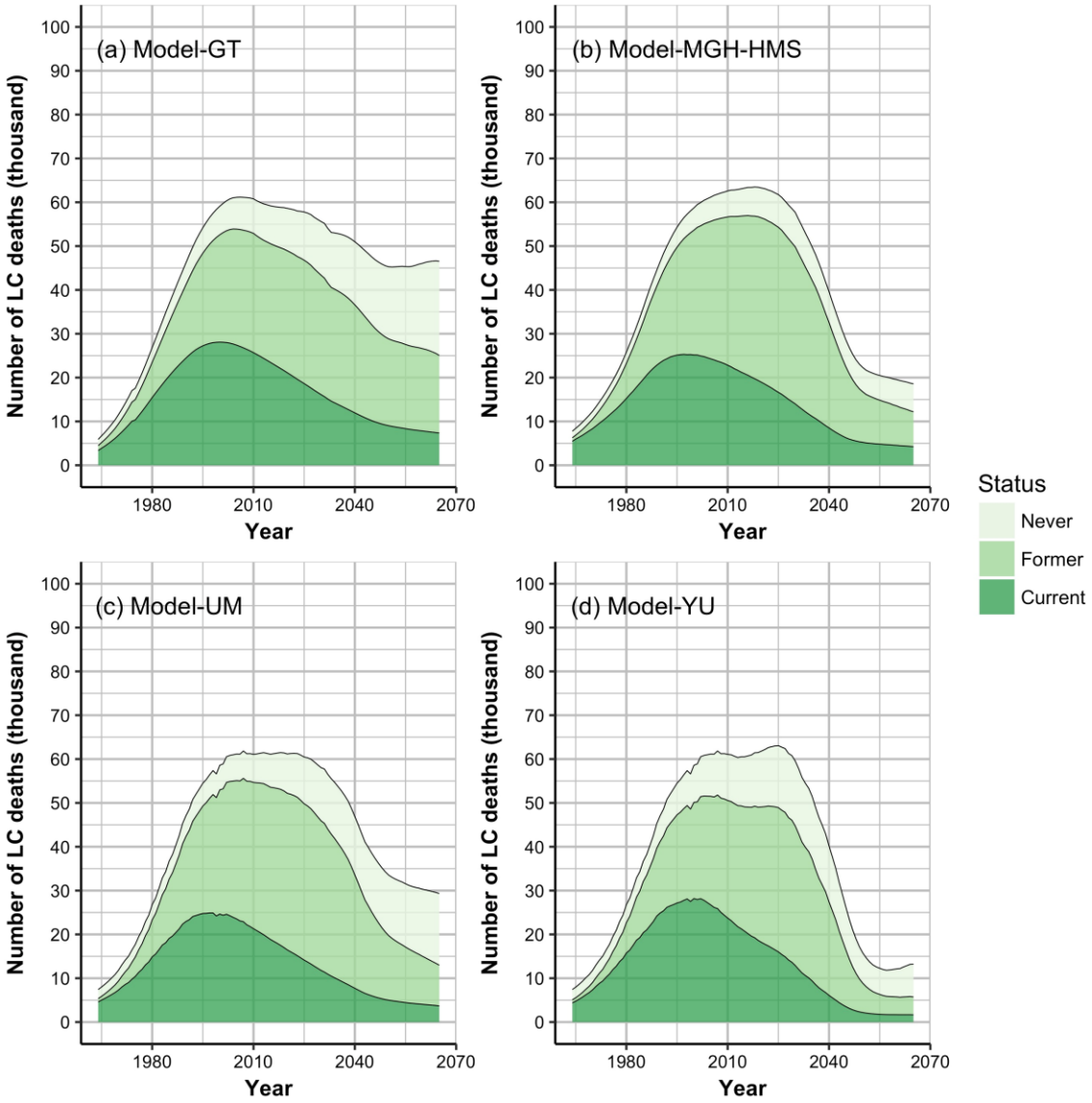
Appendix Figure 10. Population and lung cancer deaths from the University of Michigan (UM) model. Number and percentage of US women aged 30-84 by smoking status for 1964-2065 (left panels) and number and percentage of lung cancer deaths by smoking status for 1964-2065 (right panels), under the status quo scenario.



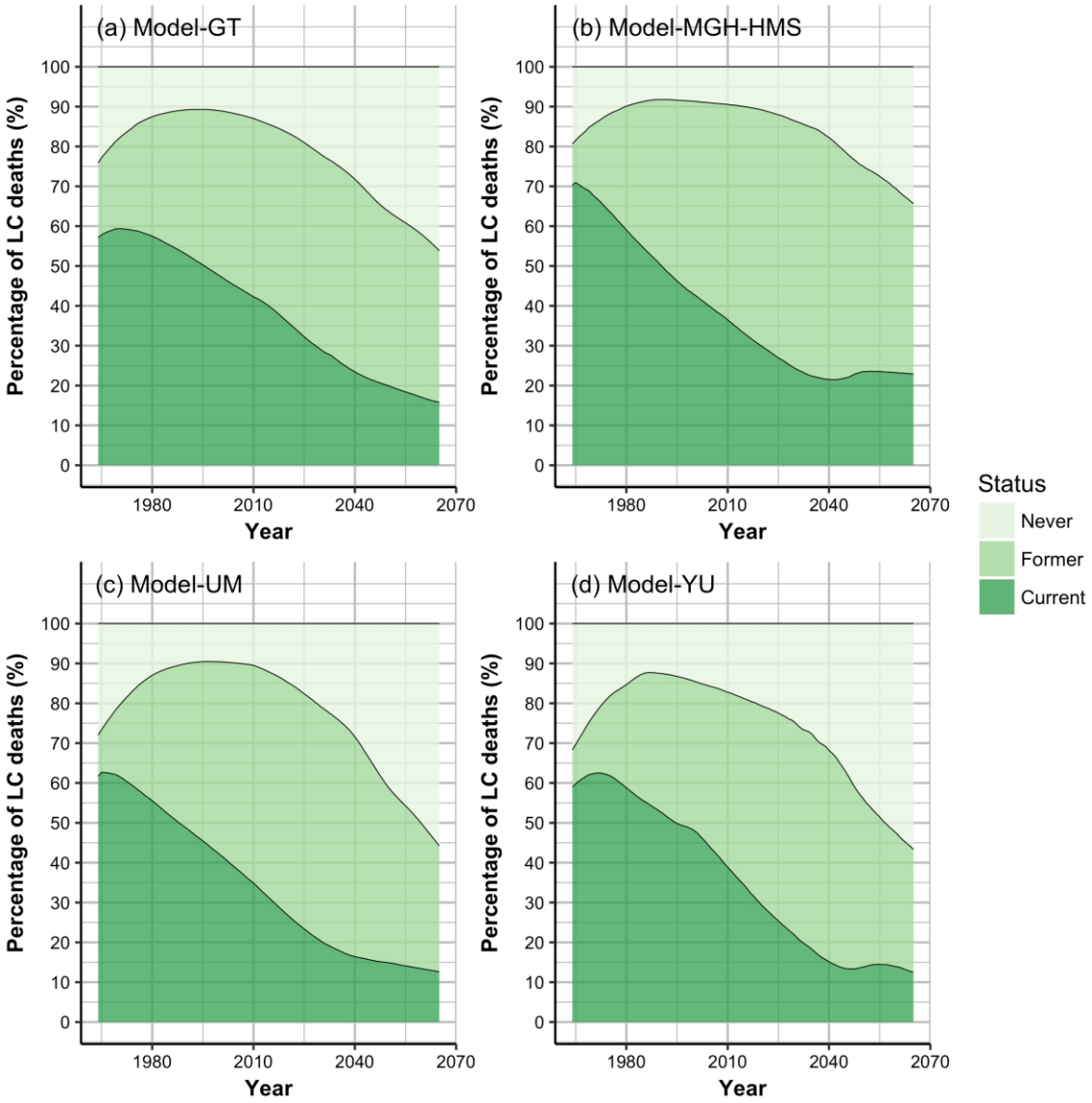
Appendix Figure 11. Number of lung cancer deaths by smoking status among US men with ages 30-84 under the status quo scenario obtained from four CISNET-Lung models: Georgetown University (GT), Massachusetts General Hospital and Harvard Medical School (MGH-HMS), University of Michigan (UM), Yale University (YU).



Appendix Figure 12. Percentage of lung cancer deaths by smoking status among US men with ages 30-84 under the status quo scenario obtained from four CISNET-Lung models: Georgetown University (GT), Massachusetts General Hospital and Harvard Medical School (MGH-HMS), University of Michigan (UM), Yale University (YU).



Appendix Figure 13. Number of lung cancer deaths by smoking status among US women with ages 30-84 under the status quo scenario obtained from four CISNET-Lung models: Georgetown University (GT), Massachusetts General Hospital and Harvard Medical School (MGH-HMS), University of Michigan (UM), Yale University (YU).



Appendix Figure 14. Percentage of lung cancer deaths by smoking status among US women with ages 30-84 under the status quo scenario obtained from four CISNET-Lung models: Georgetown University (GT), Massachusetts General Hospital and Harvard Medical School (MGH-HMS), University of Michigan (UM), Yale University (YU).

Appendix Table 1. Mean age-adjusted lung cancer mortality rates per 100,000 and number of lung cancer deaths in thousands across four CISNET-Lung models for the US population aged 30-84 under three different smoking scenarios: Status quo, optimistic and pessimistic scenarios. The 2000 US population was used as the standard to calculate age-adjusted rates. The range in parenthesis represents the results under the optimistic and pessimistic scenarios.

	Age-adjusted lung cancer mortality rates/100,000			Number of lung cancer deaths (thousands)		
Years	Men	Women	Both	Men	Women	Both
2025	51.6 (51.6,51.6)	40.9 (40.8,40.9)	45.7 (45.6,45.7)	64.8 (64.8,64.8)	60.8 (60.7,60.8)	125.6 (125.5,125.6)
2030	40.8 (40.7,40.9)	34.8 (34.7,34.9)	37.4 (37.3,37.5)	57.4 (57.4,57.5)	57.7 (57.7,57.8)	115.1 (115.0,115.3)
2035	32.2 (32.1,32.4)	29.6 (29.5,29.8)	30.7 (30.6,30.9)	48.2 (48.1,48.4)	52.0 (51.9,52.2)	100.2 (99.9,100.6)
2040	25.0 (24.7,25.3)	24.6 (24.3,24.8)	24.7 (24.4,25.0)	38.7 (38.4,39.1)	44.4 (44.1,44.7)	83.1 (82.5,83.8)
2045	19.6 (19.1,20.1)	20.0 (19.6,20.5)	19.8 (19.3,20.3)	29.9 (29.3,30.6)	35.2 (34.7,35.8)	65.2 (64.1,66.4)
2050	16.7 (16.0,17.6)	17.2 (16.6,17.9)	16.9 (16.3,17.8)	25.3 (24.3,26.6)	29.2 (28.4,30.3)	54.5 (52.7,56.8)
2055	15.5 (14.4,16.8)	15.9 (15.0,17.0)	15.7 (14.7,16.9)	24.2 (22.6,26.3)	27.4 (26.0,29.2)	51.6 (48.6,55.4)
2060	14.4 (12.9,16.3)	15.0 (13.8,16.6)	14.7 (13.3,16.5)	23.8 (21.4,26.9)	27.0 (25.0,29.7)	50.8 (46.4,56.6)
2065	13.4 (11.6,15.8)	14.4 (12.9,16.3)	13.9 (12.2,16.0)	23.4 (20.3,27.6)	26.9 (24.2,30.6)	50.4 (44.5,58.2)

Materials and Methods

Projection of smoking pattern under the status quo scenario

Under the status quo scenario, we assumed that the current smoking patterns continue into the future. For years not covered by the data, smoking initiation probabilities for ages 8 and older are obtained using the estimated parameters from the corresponding age-period-cohort model, holding period effects fixed at the estimated level for the 2015 for subsequent years and cohort effects at the estimated level for the 1997 birth cohort. Analogously, cessation probabilities and smoking intensity for ages 15 and older are kept at the value estimated for the 1985 birth cohort. The implication of extrapolating the 'status quo' projections in this way is that cross-sectional estimates of current smoking prevalence will continue to decline until everyone born before 2015 is deceased, but this decline will slow down substantially as those born before 2015 represent progressively smaller portions of the US population (Figure 1 in the main text). Under this operational definition, status quo represents a future where patterns of smoking initiation, cessation and intensity through 2015 will continue into the future. It does not, however, take into account any unrealized potential of recent tobacco control efforts, or any potential back-sliding if efforts are not continued.

Model calibration adjusting for temporal factors

Each CISNET-Lung model developed its own dose-response module in order to compute an age-specific lung cancer incidence or mortality given simulated US smoking data from the SHG by sex and birth cohort. Since these dose-response modules were built based on a specific cohort or registry data, it could not entirely capture lung cancer incidence or mortality pattern in the general US population. Therefore, each model made further adjustment for other temporal factors such as birth cohort and/or period. The cohort and/or period effects capture the influence of factors that affect lung cancer incidence or mortality independently of smoking, age and sex, such as changes in exposures to other relevant risk factors, improvements in treatment, or the impact of other preventive interventions. Two models (UM, YU) adjusted for both period and birth cohort, and the other two models (GT, MGH-HMS) only for birth cohort. More details are shown in the individual model descriptions below.

Individual Models

Georgetown University (GT) model

The Georgetown model is a macro level model that uses population data by age and sex Lung cancer mortality is first calculated by using the two-stage clonal expansion model developed based on CPS-II data for the role of smoking duration, intensity and time since quit for former smokers. Separate estimates of predicted lung cancer rates by age and sex are developed for current, former and never smokers, and then aggregated over the smoking groups for each age and sex. To correct for variations by cohort and age (where the cohort effects reflect temporal effects), regression models are then estimated

that allow for deviations in the predictions of lung cancer death rates from historical lung cancer death rates. A specification which allowed for an interaction of the predicted lung cancer rates with age and age-squared and cohort and cohort-squared along with independent cohort and cohort-squared variables best fit the data on lung cancer death rates, and were used as our final models. The model then projected the future cohort effects based on the estimates from the observed data.

Methods used to predict lung cancer death rates by smoking-related factors

To incorporate the role of smoking duration and intensity in lung cancer mortality, we used the two-stage clonal expansion (TSCE) model as applied by Hazelton et al. (1). They estimated a series of non-linear equations that related the lung cancer death rate to rates of initiation, cell division/apoptosis of initiated cells, and malignant conversion of initiated cells, which, in turn, were a function of smoking intensity and duration. Separate models were developed using CPS-I and CPS-II data, and the model based on the CPS-II data was used for the current study.

Data from Smoking History Generator (SHG) on smoking intensity, age of initiation and years since quit by age, sex, and year were applied to the TSCE models to determine lung cancer death rates separately for never, current, and former smokers. The population by smoking status were derived by Holford et al. (2–4) using age-period-cohort models applied to National Health Interview Survey data from 1965 to 2015. For current and former smokers, intensity was measured as the mean number of cigarettes smoked per day. Smoking duration was measured as the current age minus the age of smoking initiation for current smokers, and as the current age minus the sum of age of smoking initiation and the number of years quit for former smokers.

Lung cancer mortality rates were estimated for each age, sex and year (1969-2010) by smoking status (current, former and never). Death rates for males and females were separately applied by smoking status, age and year to the population in the respective categories (measured by prevalence by smoking status multiplied by the total population) to obtain total deaths. The deaths were summed over the 3 smoking status categories to obtain predicted total lung cancer deaths by age, sex and year. They were then divided by the relevant population to obtain overall lung cancer mortality rates. Historical lung cancer rates by age, sex and year were also obtained from the National Center for Health Statistics.

Comparison of Predicted to Historical Rates

The TSCE used CPS-II data, which is not representative of the U.S. population and may be subject to misclassification error. Consequently, the predictions may be biased for the population at large, and the extent of bias may vary over time if smoking or non-smoking risks (e.g. air pollution, radon, or second hand smoke) vary over time.

To detect and correct for potential biases, we estimated regression models that allow for deviations in the predictions of lung cancer death rates from historical lung cancer death rates. These models were estimated using data for each age 30 to 84 for each of the years from 1969 to 2000 to calibrate the model. We then validated the model by comparing predicted estimates by age and sex to actual lung cancer rates for the years 2001 through 2010.

We began with a simple model that regressed the historical lung cancer rate (HLCR) on the predicted lung cancer rates (PLCR) and intercept, which showed clear indication of a systematic error related to age and year. Since age, period and cohort are collinear, we focused on age and cohort, where age is known to affect lung cancer rates (5,6), and cohort to distinguish if smoking risks vary over time. We considered age, age^2 and age^3 and cohort, cohort^2 and cohort^3 sequentially to the equations, and found that the cubed term added little explanatory power and did not reduce correlation of the error terms. The resulting basic equation where a=each individual age and t=year was:

$$HLCR_{a,t} = b_0 + [b_1 + b_2 Age_{a,t} + b_3 Age_{a,t}^2 + b_4 Cohort_{a,t} + b_5 Cohort_{a,t}^2]\ PLCR_{a,t} + e_{a,t,} \quad (1)$$

The first term in brackets (b_1) indicates general biases in the TSCE estimates (unrelated to age or year), the next two terms correct the predictions of the smoking models for linear(b_2) and non-linear(b_3) biases in age, followed by linear(b_4) and non-linear(b_5) biases by cohort. The cohort coefficients were used to consider the changing relationship of smoking to lung cancer over time.

We next estimated equations with the cohort terms grouped. Out of the 85 cohorts observed from 1969 to 2000, we also distinguished cohort sub-groups (1-5, 6-14, 15-24, 25-44, 45-54, 55-64, and 65-85, because of the similar pattern of error terms observed within those cohort groups (based on error terms in the equations with age variables).

$$HLCR_{a,t} = b_0 + [b_1 + b_2 Age_a + b_3 Age_a^2 + b_4 Cohort(1\text{-}5)_{a,t} + b_5 Cohort\ (15\text{-}24) + b_6 * Cohort(25\text{-}44)_{a,t} + b_7 * Cohort(45\text{-}54)_{a,t} + b_8 * Cohort(55\text{-}64)_{a,t} + b_9 * Cohort(65+)_{a,t}] * PLCR + e_{a,t.} \quad (2)$$

The coefficients have similar interpretations to equation (1).

For each of the above models, we also considered the effect of non-smoking related age and cohort effects by including non-interacted age and cohort quadratic terms in the equations. We found that the cohort-terms alone provided less biased predictions over the years 2001-2010 than the age terms alone, and that multicollinearity was induced when age terms were added to the cohort terms. As an alternative method for considering changes in the smoking and non-smoking risks over time, we estimated models with age and period effects, where the period effects were intended to capture changing risks over time. The results implied similar variation in risk over time as those in the cohort model, but performed less well in terms of validation with actual lung cancer rates, R^2 and systematic bias in the error terms.

Massachusetts General Hospital and Harvard Medical School (MGH-HMS) model

The Massachusetts General Hospital and Harvard Medical School (MGH-HMS) group's model is a microsimulation model which simulates an individual patient's lung cancer development, progression, detection, treatment, and survival (7,8). The MGH-HMS model was developed to evaluate the clinical effectiveness and cost-effectiveness of low-dose computed tomography (CT) screening for lung cancer (9–13). The model has also been used to estimate the impact of reduced tobacco smoking on lung cancer mortality in the United States (14).

The MGH-HMS model initially populates with disease-free individuals who then go through different health states according to monthly transition probabilities. In each monthly cycle, an individual may develop lung cancer, have an existing cancer grow, or develop symptoms or metastases. The risk of lung cancer is related to individuals' smoking history, which is updated monthly (the model also includes cancers in non-smokers). Smoking exposure is supplied by the Smoking History Generator, a module developed by CISNET(15). Lung cancers can be detected by an evaluation of symptoms, through incidental imaging, or by CT screening (with different tumor behavior for screen-detected cases). Individuals with suspected lung cancer receive diagnostic and staging tests, and then may undergo treatment. The screening module can be turned on or off to allow for analyses of treatment effectiveness for screen vs. non-screen detected cases.

Each hypothetical individual in the MGH-HMS model can develop up to three cancers from any of five lung cancer cell types (adenocarcinoma, large cell, squamous cell, small cell, and other). For each cell type, the monthly probability of cancer development is described by a logistic equation with seven natural history parameters including a type-specific intercept, type-specific coefficients for age, age^2, years of cigarette exposure (smoke-years, SY), an interaction term between SY and age^2, the mean number of cigarettes smoked per day (cigarettes per day, CPD), and the years since quitting (YSQ) smoking. The natural history parameters related to unobservable events (i.e. the initiation of the first cancer cell) were estimated by calibrating the model using SEER registry data (cancer incidence by cell type, stage distribution at diagnosis, and stage-specific survival), published cohort studies, and clinical trial data. The details of model calibration and validation of the original natural history parameters have been described in our previous publications (7,16).

For this analysis, we used an age-cohort formalism to further capture the lung cancer mortality rates observed in the United States. To account for the changes in unmeasured risk factors (in addition to the change in smoking pattern) experienced by different birth cohorts, we multiplied a sex-specific cohort coefficient, b_{BY}, by the monthly probability of lung cancer development. The age-dependence of lung cancer mortality rate for the males born in 1930, our reference birth cohort, is driven by the logistic equation and the natural history parameters described above. For other birth cohorts born between years 1900 and 1970, the values of b_{BY} are determined by calibrating the model outputs to the observed lung cancer mortality rates stratified by 5-year birth cohort groups. The calibration of b_{BY} was done using lung cancer mortality data prior to calendar year 2000. Using the model to extrapolate mortality rate beyond calendar year 2000, the values of b_{BY} are fixed to the value of last birth year. In the literature, age-period-cohort method is often used to analyze cancer incidence or mortality (17,18). However, we decided not to include a period coefficient in our model to avoid overfitting the results and difficulties in extrapolating the model outputs beyond calendar year 2000.

University of Michigan (UM) model

The University of Michigan model was developed within the framework of the Two-Stage clonal expansion (TSCE) model, which represents the process of carcinogenesis in three phases. In the first phase (initiation), a susceptible stem cell acquires one or more mutations resulting in an initiated cell, which has partially escaped growth control. In the second phase (promotion), initiated cells undergo clonal expansion, either

spontaneously or in response to endogenous or exogenous promoters. Finally, in the third phase (malignant conversion), one of the initiated cells acquires further mutational changes leading to a malignant cell.

Smoking dose-response module

The TSCE model and its extensions have been used for the analyses of various cancer sites, including lung(1,19), colon (20–22), esophagus (23), and pancreas (21).
For this study, we used the TSCE models built on lung cancer mortality data in Nurses' Health Study (NHS, 1976-2008) for women and Health Professionals' Follow-up Study (HPFS, 1986-2008) for men. To model the effects of smoking on lung cancer risk, the model initiation, promotion, and malignant transformation parameters are assumed to alter during periods of smoking exposure through flexible dose-response relationships:

$$\theta(t) = \theta_0 \times \left(1 + \theta_1 \times d(t)^{\theta_2}\right),$$

where θ represents identifiable biological parameters in the TSCE model, θ_0 the background rate, θ_1 the dose-response coefficient, θ_2 the non-linearity of the dose-response, and d(t) smoking dose at time t. This dose-response relationship links the individual smoking history to the cell kinetic parameters in the TSCE model. Appendix Tables 2-3 present the TSCE model structure and estimated parameters.

Appendix Table 2. Model parameters for background rate and smoking dose–response relationship

Background variables	
$X=10^7$	Assume 10^7 normal stem cells in both lungs
α_0	Background cell division rate (per cell per year)
$g_0=\alpha_0-\beta_0-\mu_0$	Background net cell promotion rate (per cell per year)
$\nu_0=\mu_0$	Background initiation rate; Background malignant transformation rate (per cell per year)
t_{lag}=5 years	Fixed constant lag time
Dose-response variables	
NHS* and HPFS† models	
$\nu_i=\nu_0$	Initiation rate (per cell per year); No dose-response
$g_i=g_0(1+p_1\times dose_i^{p2})$	Net initiated cell promotion rate (per cell per year)
$\alpha_i=\alpha_0(1+p_1\times dose_i^{p2})$	Initiated cell division rate (per cell per year)
$\mu_i=\mu_0(1+p_3\times dose_i^{p4})$	Malignant transformation rate (per cell per year)

* Nurses' Health Study; † Health Professionals Follow-Up Study

Appendix Table 3. Parameter estimates for the NHS and the HPFS models

Model	α_0	g_0	ν_0 (=μ_0)	p_1	p_2	p_3	p_4
Lung Cancer mortality models							
NHS*	3.00	0.076	1.03×10^{-7}	0.20	0.50	0.05	0.60
HPFS†	3.00	0.076	1.03×10^{-7}	0.33	0.35	0.21	0.18

* Nurses' Health Study; † Health Professionals Follow-Up Study

Prediction of lung cancer mortality in the US

We used the smoking history generator (SHG) to simulate the entire US population with detailed individual level smoking histories from 1964-2065. Then we computed age-

specific lung cancer mortality by using the TSCE models built on the NHS and the HPFS studies. These studies, however, may not represent the general US population. In addition, our models do not incorporate other risk factors such as second hand smoke, exposure to radon gas, asbestos or other carcinogens, family history, chronic obstructive pulmonary disease, occupational exposure, and race, and socioeconomic status. As a result, a cohort-specific TSCE model does not predict lung cancer mortality in the US completely, and further calibration by adjusting for secular temporal trends was necessary; we used age-period-cohort (APC) models (24) for this purpose. Standard methods of Poisson regression were used to estimate the 1-year period and 5-year birth cohort group effects in the APC models by fitting to the observed US lung cancer mortality data from 1969-2010.

To project lung cancer mortality in the US over 1964-2065, we extrapolated 1969-2010 period effects both backward (1964-1968) and forward (2011-2065) and also birth cohort effects accordingly. The UM model assumed that the cohort effect at birth year 1980 remains the same for future birth cohorts, but extrapolated the period effects differently for current, former and never smokers because other interventions, such as low-dose computed tomography (CT) lung cancer screening and improvements in treatment, could affect lung cancer mortality trends differentially by smoking status. The model applied a trend attenuation approach, the Nordpred method (25,26), to extrapolate the period effects for current and former smokers in future years, but fixed at the value for the period 2010 for never smokers.

Following the Nordpred method, the period effects for current and former smokers in future years were extrapolated out to eleven 5-year interval periods (2011-2015, 2016-2020, 2021-2025,…,2061-2065). The linear drift estimated from the most recent 5-year period effects (2006-2010) was attenuated by 20%, 40%, 60%, and 80% for the first (2011-2015), second (2016-2020), third (2021-2025), and fourth (2026-2030) 5-year projection periods, respectively, and 100% for the remaining 5-year projection periods. A similar extrapolation scheme for the cohort and period effects was used for the model validation, and this approach provided good prediction on lung cancer mortality in future years. As an example, the Appendix Figure 6 shows the lung cancer mortality projection for years 2001-2010, using each group model calibrated to the US lung cancer mortality data from 1969-2000.

Similarly, for the extrapolation of the period effects for 1964-1968, we assumed the same 5-year linear trend estimated from the period effects for 1969-1973.

Yale University (YU) model

The Yale Lung Cancer Model describes the impact of the distribution of exposure histories for cigarette smoking for cohorts of individuals as they grow older. It makes use of (a) two-stage clonal expansion (TSCE) model of carcinogenesis which describes the quantitative relationship between smoking history and lung cancer mortality (27), (b) distribution of smoking history summaries, and (c) calibration that adjusts for discrepancies between mortality probabilities derived from (a) and (b) and observed mortality rates in the US population. Let Z(a,c) represent a summary of smoking history for individuals age *a* in cohort *c*, and $I_{+}(Z(a,c))$ the overall lung cancer mortality rate estimated using TSCE and the estimated summary of smoking history for (a,c).

Calibration of the rate is accomplished by introducing an estimated multiplicative factor that may either be a constant or a function of parameters that can depend on times from critical reference points, giving rise to an estimated calibrated rate for the population,

$$I_+(a,c)^* = q(a,p,c)I_+(Z(a,c)),$$

where $q(a,p,c)$ represents the calibration factor, which depends on age and cohort, as well as, period, $p = a + c$.

Smoking history parameters

Parameters used to characterize smoking history are based on a compartment model in which a subject begins to smoke at some point after which they may quit. While this over simplifies a process that can be much more complex in reality, it does provide a useful characterization of the experience for most of the population. Summary parameters for smoking history in the US were derived from the National Health Interview Surveys (NHIS) and details are described by Holford et al.(3). Histories for the population are summarized by estimates of the conditional probability of smoking initiation and cessation, prevalence of never, current and former smokers, and the distribution of smoking intensity. Smoking intensity is given by a discrete distribution, $g(k \mid a,c)$ for the cigarette per day categories (CPD) with approximate mean in the parenthesis:

1. CPD≤5 (3)
2. 5< CPD≤15 (10)
3. 15<CPD≤25 (20)
4. 25<CPD≤35 (30)
5. 35<CPD≤45 (40)
6. 45<CPD (60).

These parameters are used in the two-stage clonal expansion model to estimate lung cancer mortality rates by single year ages and by birth cohorts. The surveys are cross-section and not longitudinal, which makes it impossible to obtain some detail on changing practices that would be desirable, and this is an unavoidable limitation.

Prevalence of ever, never, current and former smokers is represented by $P_E(a,c)$, $P_N(a,c)$, $P_C(a,c)$ and $P_F(a,c)$, respectively. The smoking initiation probability, $p(a,c)$, is the conditional probability of smoking initiation at age *a* for cohort *c*, given not a smoker at a-1, i.e.,

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

It is related to the cumulative proportion of ever smokers at age *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 $P_E(a,c) \leq P_E^*(a,c)$. The

relationship between cross-sectional and cumulative prevalence of ever smokers is given by the ratio, $C(a,c) = P_E(a,c) / P_E^*(a,c) \leq 1$. Smoking cessation is assumed to be a function of age for each cohort. The smoking cessation probability conditional on the subject being alive and currently smoking is

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

We assumed that $q(a,c) = 0$ for $a < 15$, and the cumulative proportion of smokers in cohort *c* who had not ceased smoking by age *a* is given by

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

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.

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.,

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

Estimate of lung cancer mortality

The TSCE model depends on parameters estimated using follow-up data from the Health Professionals' Follow-up Study (HPFS) for males and the Nurses' Health Study (NHS) for females. Moolgavkar et al. (1,20,28,29) proposed the TSCE model in which the carcinogenesis process is initiated in a cell that multiplies to form a clone of cells initiated for risk of developing into cancer cells (27). A second hit on one of these initiated cells transform it into a cancer cell that subsequently multiplies further until it forms a tissue mass that can be clinically identified as cancer. The functional form for the TSCE model is complex, but it has been found to provide an excellent description of the effect of age on lung cancer incidence and mortality.

To model the effect of smoking on lung cancer mortality rates, we regard the population as a mixture of never (N), current (C) and former (F) smokers, each with prevalence $P_N(a,c)$, $P_C(a,c)$ and $P_F(a,c)$ respectively, giving the overall rate

$$\lambda_+(a,c) = P_N(a,c)\lambda_N(a,c) + P_C(a,c)\overline{\lambda_C}(a,c) + P_F(a,c)\overline{\lambda_F}(a,c), \qquad (3)$$

where $\lambda_N(\cdot)$ is the rate for never smokers, $\overline{\lambda_C}(\cdot)$ and $\overline{\lambda_F}(\cdot)$ are the rates for the corresponding smoking categories of smokers, taking into account the distributions of

exposure ages and intensities. These models depend on age (*a*), age of smoking initiation (a_I), age quit (a_Q), and number of cigarettes smoked per day (*d*).

Among those who never smoked, the mortality rate $\lambda_N(a)$, is a function of age alone which reflects the underlying effect of the aging process on lung cancer risk.

For current smokers age *a* in cohort *c*, the distribution function for age of initiation, a_I, is $F_I(a_I \mid a,c)$, where $F_I(0 \mid a,c) = 0$ and

$$F_I(a_I \mid a,c) = 1 - \left[1 - F_I(a_I - 1 \mid a,c)\right]\left[1 - p(a_I - 1 \mid a,c)\right]$$

for $0 < a_I \le a$. This may be used to determine the probability density of initiation ages

$$f_I(a_I \mid a,c) = \frac{F_I(a_I \mid a,c) - F_I(a_I - 1 \mid a,c)}{F_I(a \mid a,c)} \quad \text{for } a_I = 1, \ldots, a.$$

The density function for smoking intensity is $g(k \mid a,c)$ for the k-th intensity level, and we assume that initiation age and smoking intensity are independent. TSCE provides estimates of the lung cancer mortality rate, $\lambda_C(a_I,k)$ for age at initiation, a_I and smoking intensity, *k*, which yields the overall rate for the mixture of exposures in the population,

$$\overline{\lambda_C}(a,c) = \sum_{i=1}^{a}\sum_{k=1}^{6} f_I(i \mid a,c)\, g(k \mid a,c)\, \lambda_I(i,k).$$

The cumulative distribution of cessation at a_Q, given initiation at age a_I, current age *a*, and cohort *c* is

$$F_Q(a_Q \mid a_I,a,c) = 1 - \left[1 - F_Q(a_Q - 1 \mid a_I,a,c)\right]\left[1 - q(a_Q - 1 \mid a_I,a,c)\right],$$

and the corresponding probability density function is

$$f_Q(a_Q \mid a_I,a,c) = \frac{F_Q(a_Q \mid a_I,a,c) - F_Q(a_Q - 1 \mid a_I,a,c)}{F_Q(a \mid a_I,a,c)} \quad \text{for } a_Q = a_I, \ldots, a \text{ and } a_I = 1, \ldots, a.$$

The chain rule yields the joint distribution of initiation and cessation times

$$f_{IQ}(a_I,a_Q \mid a,c) = f_Q(a_Q \mid a_I,a,c) f_I(a_I \mid a,c).$$

Using the estimate of this joint distribution, the overall rate for former smokers at age *a* and cohort *c* is as following:

$$\overline{\lambda_F}(a,c) = \sum_{i=1}^{a}\sum_{j=1}^{a}\sum_{k=1}^{6} f_Q(j \mid i,a,c) f_I(i \mid a,c) g(k \mid a,c) \lambda_{IQ}(i,j,k).$$

Hence, we have estimates of the lung cancer mortality rate for never, current and former smokers, which are required to estimating the overall mortality rate at age *a* and cohort *c* from the TSCE model in equation (3).

Calibration and validation

An age-period-cohort (APC) model is employed to calibrate the carcinogenesis model in order to bring rates into conformity with rates for the overall population. Let $t = (a,p,c)$ represent a vector of temporal elements: age, period and cohort, respectively. Smoking

history for the population is represented by $Z(t)$, which depends on the estimated smoking histories obtained from NHIS surveys and the TSCE model, $l\{Z(t)\}$. Calibrated estimates are determined by estimating a multiplicative factor that depends on the temporal vector,

$$l^*\{Z(t);t\} = q(t)\, l\{Z(t)\}, \quad (4)$$

which is a log-linear function of the temporal elements, similar to the approach employed by Meza et al. (19),

$$q(a,p,c) = \exp\{m + a_a + p_p + g_c\}. \quad (5)$$

The intercept, m, scales the rates so that the estimates from the model correspond overall with those observed in the US population. Estimates of temporal elements for age $(\alpha_a : a = 1,...,A)$, period $(\pi_p : p = 1,...,P)$, and cohort $(g_c : c = 1,...,C)$ provide the estimated calibration factor. If temporal effects are all 0, then the model is in good temporal agreement with the population, and the extent to which these effects become parallel to the abscissa indicates the adequacy of the carcinogenesis model and the estimates of exposure histories to characterize temporal trend in the population rates. Poor agreement could result from either a limitation in the carcinogenesis model or an inaccurate estimate of exposure history for the population.

The well recognized identifiability problem in APC models also applies to estimates of parameters in the calibration function, and the phenomenon has been discussed in considerable detail previously (30–34). In this form, $\log q$ resembles an analysis of variance model, and the usual constraints imply that

$$\sum_a a_a = \sum_p p_p = \sum_c g_c = 0,$$

but the linear dependence among age, period, and cohort extends to indices for the three time effects, in that $c = p - a + A$. Hence, the design matrix for a linear model that includes all three factors is not of full rank, and a unique set of parameters for a corresponding generalized linear model does not exist (30,31). The primary analysis assumes that the TSCE model accurately characterized the effect of age, *a*, so that $a_a = 0$ for all *a*.

Calibration requires fitting the APC model for $q(x)$ to a function of the observed rates, and thus obtaining optimal estimates of the temporal parameters for calibration. We assume that the number of lung cancer deaths, *Y*, has a Poisson distribution, and the denominator for the rate, *D*, is known. The observed calibration factor, $\hat{q} = Y / Dl$, is the maximum likelihood estimate for the group, and the variance of the estimate would be $Var(\hat{q}) = q / Dl$. If we also assume a log-linear model for the calibration factor, then maximum likelihood estimates of the parameters can be obtained by fitting a generalized linear model in which the linear predictor, h, is related to the calibrated rate, l^* through the link function

$$h = \log q = \log\left(l^* / l\right) = m + a_a + p_p + g_c.$$

We specify a Poisson distribution for the response (i.e., the observed calibration factor) and introduce a scale weight equal to the denominator for the factor, Dl (35–37). Estimates of the model parameters were obtained using PROC GENMOD in SAS.

Estimates of a calibrated rate given a particular set of smoking exposure covariates, *Z*, employs both the estimated rate from the carcinogenesis model and the corresponding maximum likelihood estimate of the calibration factor for the given age, period, and cohort, $\hat{q}(a,p,c) / \{Z(a,p,c)\}$.

The observed mortality data in single-year age and period were used to calibrate the model beginning in 1969 and ending in 2010 (2000 in the case of the validation study). In order to project estimates to the range of interest, 1964-2065, it is necessary to extend the period and cohort effects from 1969 to 1964 and 2010 to 2065 (2000 to 2010 for the validation study). For the period effects, the linear trend estimated from periods 2005-2010 was applied to extrapolate the effects for 2011-2013, and the period effect was held at the 2013 level afterwards. The model extrapolated the cohort effects for 1991-2065 by applying the average of the parameters estimates for 1981-1990.

References:

1. Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. Cancer Epidemiol Biomarkers Prev. 2005 May;14(5):1171–81.

2. Anderson CM, Burns DM, Dodd KW, Feuer EJ. Chapter 2: Birth-cohort-specific estimates of smoking behaviors for the U.S. population. Risk Anal. 2012 Jul;32 Suppl 1:S14-24.

3. Holford TR, Levy DT, McKay LA, Clarke L, Racine B, Meza R, et al. Patterns of birth cohort-specific smoking histories, 1965-2009. Am J Prev Med. 2014 Feb;46(2):e31-7.

4. Holford TR, Meza R, Warner KE, Meernik C, Jeon J, Moolgavkar SH, et al. Tobacco control and the reduction in smoking-related premature deaths in the United States, 1964-2012. JAMA. 2014 Jan 8;311(2):164–71.

5. Flanders WD, Lally CA, Zhu BP, Henley SJ, Thun MJ. Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: results from

Cancer Prevention Study II. Cancer Res. 2003 Oct;63(19):6556–62.

6. Knoke JD, Shanks TG, Vaughn JW, Thun MJ, Burns DM. Lung cancer mortality is related to age in addition to duration and intensity of cigarette smoking: an analysis of CPS-I data. Cancer Epidemiol Biomarkers Prev. 2004 Jun;13(6):949–57.

7. Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. Value Health. 2009 Jun;12(4):521–9.

8. McMahon PM, Kong CY, Johnson BE, Weinstein MC, Weeks JC, Tramontano AC, et al. Chapter 9: The MGH-HMS lung cancer policy model: tobacco control versus screening. Risk Anal. 2012 Jul;32 Suppl 1:S117-24.

9. McMahon PM, Kong CY, Johnson BE, Weinstein MC, Weeks JC, Kuntz KM, et al. Estimating long-term effectiveness of lung cancer screening in the Mayo CT screening study. Radiology. 2008 Jul;248(1):278–87.

10. McMahon PM, Kong CY, Weinstein MC, Tramontano AC, Cipriano LE, Johnson BE, et al. Adopting helical CT screening for lung cancer: potential health consequences during a 15-year period. Cancer. 2008 Dec;113(12):3440–9.

11. McMahon PM, Kong CY, Bouzan C, Weinstein MC, Cipriano LE, Tramontano AC, et al. Cost-effectiveness of computed tomography screening for lung cancer in the United States. J Thorac Oncol. 2011 Nov;6(11):1841–8.

12. De Koning HJ, Meza R, Plevritis SK, Ten Haaf K, Munshi VN, Jeon J, et al. Benefits and harms of computed tomography lung cancer screening strategies: A comparative modeling study for the U.S. Preventive services task force. Ann Intern Med. 2014;160(5).

13. McMahon PM, Meza R, Plevritis SK, Black WC, Tammemagi CM, Erdogan A, et

al. Comparing benefits from many possible computed tomography lung cancer screening programs: Extrapolating from the National Lung Screening Trial using comparative modeling. PLoS One. 2014;9(6).

14. Moolgavkar SH, Holford TR, Levy DT, Kong CY, Foy M, Clarke L, et al. Impact of reduced tobacco smoking on lung cancer mortality in the United States during 1975-2000. J Natl Cancer Inst. 2012 Apr 4;104(7):541–8.

15. Jeon J, Meza R, Krapcho M, Clarke LD, Byrne J, Levy DT. Chapter 5: Actual and counterfactual smoking prevalence rates in the U.S. population via microsimulation. Risk Anal. 2012 Jul;32 Suppl 1:S51-68.

16. Meza R, Ten Haaf K, Kong CY, Erdogan A, Black WC, Tammemagi MC, et al. Comparative analysis of 5 lung cancer natural history and screening models that reproduce outcomes of the NLST and PLCO trials. Cancer. 2014;120(11).

17. Zheng T, Holford TR, Chen Y, Ma JZ, Mayne ST, Liu W, et al. Time trend and age-period-cohort effect on incidence of bladder cancer in Connecticut, 1935-1992. Int J cancer. 1996 Oct;68(2):172–6.

18. Holford TR. Approaches to fitting age-period-cohort models with unequal intervals. Stat Med. 2006 Mar;25(6):977–93.

19. 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 Apr;19(3):317–28.

20. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. Proc Natl Acad Sci U S A. 2002 Nov;99(23):15095–100.

21. Meza R, Jeon J, Moolgavkar SH, Georg Luebeck E. Age-specific incidence of

cancer: Phases, transitions, and biological implications. Proc Natl Acad Sci U S A. 2008;105(42).

22. Meza R, Jeon J, Renehan AG, Luebeck EG. Colorectal cancer incidence trends in the United States and United Kingdom: Evidence of right- to left-sided biological gradients with implications for screening. Cancer Res. 2010;70(13).

23. 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).

24. Holford TR. Understanding the Effects of Age, Period, and Cohort on Incidence and Mortality Rates. Annu Rev Public Health [Internet]. 1991 May [cited 2018 Jul 16];12(1):425–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2049144

25. Moller B, Fekjaer H, Hakulinen T, Sigvaldason H, Storm HH, Talback M, et al. Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. Stat Med. 2003 Sep;22(17):2751–66.

26. Virani S, Sriplung H, Rozek LS, Meza R. Escalating burden of breast cancer in southern Thailand: analysis of 1990-2010 incidence and prediction of future trends. Cancer Epidemiol. 2014 Jun;38(3):235–43.

27. Hazelton WD, Jeon J, Meza R, Moolgavkar SH. Chapter 8: The fhcrc lung cancer model. Risk Anal. 2012;32(SUPPL.1).

28. Moolgavkar SH, Dewanji A, Venzon DJ. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. Risk Anal. 1988 Sep;8(3):383–92.

29. Moolgavkar SH, Luebeck G. Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. Risk Anal. 1990 Jun;10(2):323–41.

30. Fienberg SE, Mason WM. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In: Schuessler KF, editor. San Francisco: Jossey-Bass, Inc; 1978. p. 1–67. (Sociological Methodology 1979).

31. Holford TR. The estimation of age, period and cohort effects for vital rates. Biometrics. 1983 Jun;39(2):311–24.

32. Kupper LL, Janis JM, Karmous A, Greenberg BG. Statistical age-period-cohort analysis: a review and critique. J Chronic Dis. 1985;38(10):811–30.

33. Kupper LL, Janis JM, Salama IA, Yoshizawa CN, Greenberg BG, Winsborough HH. Age-period-cohort analysis: an illustration of the problems in assessing interaction in one observation per cell data. Commun Stat - Theory Methods [Internet]. 1983 Jan 27 [cited 2018 Aug 17];12(23):201–17. Available from: http://www.tandfonline.com/doi/abs/10.1080/03610928308828640

34. Holford TR. Age-period-cohort analysis. In: Armitage P, Colton T, editors. Chichester: John Wiley & Sons; 1998. p. 82–99. (Encyclopedia of Biostatistics).

35. McCullagh P, Nelder JA. Generalized Linear Models. Second. London: Chapman and Hall; 1989.

36. Aranda-Ordaz FJ. On two families of transformations to additivity for binary response data [Internet]. Vol. 68, Biometrika. 1981 [cited 2018 Aug 17]. Available from: https://academic.oup.com/biomet/article-abstract/68/2/357/260360

37. Holford TR. Multivariate Methods in Epidemiology [Internet]. Oxford University Press; 2002 [cited 2018 Aug 17]. Available from: http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195124408.001.0001/acprof-9780195124408