Dynamics of the Risk of Smoking-Induced Lung Cancer

A Compartmental Hidden Markov Model for Longitudinal Analysis

Marc Chadeau-Hyam,[a] *Pascale Tubert-Bitter,*[b,c] *Chantal Guihenneuc-Jouyaux,*[d] *Gianluca Campanella,*[a] *Sylvia Richardson,*[e] *Roel Vermeulen,*[f] *Maria De Iorio,*[g] *Sandro Galea,*[h] *and Paolo Vineis*[a]

Background: To account for the dynamic aspects of carcinogenesis, we propose a compartmental hidden Markov model in which each person is healthy, asymptomatically affected, diagnosed, or deceased. Our model is illustrated using the example of smoking-induced lung cancer.

Methods: The model was fitted on a case-control study nested in the European Prospective Investigation into Cancer and Nutrition study, including 757 incident cases and 1524 matched controls. Estimation was done through a Markov Chain Monte Carlo algorithm, and simulations based on the posterior estimates of the parameters were used to provide measures of model fit. We performed sensitivity analyses to assess robustness of our findings.

Results: After adjusting for its impact on exposure duration, age was not found to independently drive the risk of lung carcinogenesis, whereas age at starting smoking in ever-smokers and time since cessation in former smokers were found to be influential. Our data did not support an age-dependent time to diagnosis. The estimated time between onset of malignancy and clinical diagnosis ranged from 2 to 4 years. Our approach yielded good performance in reconstructing individual trajectories in both cases (sensitivity >90%) and controls (sensitivity >80%).

Conclusion: Our compartmental model enabled us to identify time-varying predictors of risk and provided us with insights into the dynamics of smoking-induced lung carcinogenesis. Its flexible and general formulation enables the future incorporation of disease states, as measured by intermediate markers, into the modeling of the natural history of cancer, suggesting a large range of applications in chronic disease epidemiology.

(*Epidemiology* 2014;25: 28–34)

Evidence is accumulating in chronic disease epidemiology to suggest that disease risk is governed not only by cumulative levels of exposure but also by dynamic aspects of its history. This has been formalized within the exposome[1–3] and life-course epidemiology[4] concepts according to which the risk of chronic disease could be better defined and subsequently predicted by characterizing the individual chemical environment—in turn, defined by the biological response to external exposures at several critical time points in life.

Several regression-based approaches already include dynamic aspects of exposure and measure their impact in risk inferences.[5,6] Novel approaches rely on the application of methods developed in infectious disease epidemiology to study chronic diseases,[7] where longitudinal models aim at the prediction of both the size and the dynamics of an epidemic, and hence by design include a temporal component in causal inferences.

Compartmental models, in which the population is subdivided into several states reflecting their health status, have been particularly successful, notably in the study of AIDS/HIV infection.[8] Health states are either observed[9,10] or hidden,[11,12] and the purpose of such methods was to estimate, based on the observed individual or population-based trajectories, the transition probabilities between (observed or hidden) states that drive the dynamics of the disease natural history.

As part of causal diagram approaches, compartmental models constitute an explicit and intuitive representation of causal structures linking exposures and outcomes.[13–15] Multistage models, developed in cancer epidemiology, have been used to infer biological pathways[16,17] and to provide insight into putative cellular mechanisms involved in

Submitted 23 September 2012; accepted 13 August 2013; posted 14 November 2013.

From the [a]MRC/HPA Centre for Environment and Health, School of Public Health, Imperial College, London, United Kingdom; [b]Centre for Research in Epidemiology and Population Health, INSERM, Villejuif, France; [c]UMRS 1018, University Paris Sud, Villejuif, France; [d]EA 4064, University Paris Descartes, Paris, France; [e]MRC Biostatistics Unit, Institute of Public Health, Cambridge, Uinted Kingdom; [f]Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands; [g]Department of Statistical Science, University College London, London, United Kingdom; and [h]Department of Epidemiology, Mailman School of Public Health, Columbia University, NY.

This work has been carried out within the Transdisciplinary Research in Cancer of the Lung (TRICL) project, which is supported by the National Cancer Institute, Grant U19 CA148127 02 to Chris Amos. Marc Chadeau-Hyam, Roel Vermeulen and Paolo Vineis acknowledge the European FP7 EnviroGenoMarkers (Grant Agreement 226756 to S.A. Kyrtopoulos) and Exposomics (Grant Agreement 308610 to P. Vineis) projects. Roel Vermeulen and Paolo Vineis acknowledge the European FP7 ECNIS 2 project (Grant Agreement 266198 to K. Rydzynski).

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the author.

Correspondence: Marc Chadeau-Hyam, St Mary's Hospital, Norfolk Place, W21PG London, UK. E-mail: m.chadeau@imperial.ac.uk.

ISSN: 1044-3983/14/2501-0028
DOI: 10.1097/EDE.0000000000000032

carcinogenesis and their potential control.[18–21] As an extension of these approaches to fit macroscopic data, and to ease the interpretability of model parameters, we propose an individual-based compartmental model for the natural history of smoking-induced lung cancer using case-control data nested in a large longitudinal prospective cohort study. Several approaches have already been proposed to infer measures of absolute risk of lung cancer from multistate models applied to case-control data set making use of external information to characterize absolute lung cancer rates in the population.[22–25] Our approach differs from these, in that it does not provide absolute risk measures but seeks for the determinants of the disease progression to ensure optimal reconstruction, at the individual level, of disease development across the full study population.

We chose a compartmental hidden Markov model derived from a previous study,[26] whose structure has been adapted to model the natural history of lung carcinogenesis at the individual level. Parameter estimation is done through a Markov Chain Monte Carlo procedure implemented in C^{++}, which is detailed in the eAppendix (http://links.lww.com/EDE/A742) and freely available on the author's Web site: http://www.imperial.ac.uk/people/m.chadeau.

METHODS

EPIC Data

The European Prospective Investigation into Cancer and Nutrition (EPIC) study[27] is a large prospective cohort study with over 520,000 volunteer subjects enrolled between 1992 and 2000 from 23 centers in 10 Western European countries.[28,29] We use data from a lung cancer case-control study nested in the cohort; 757 incident cases and 1524 controls were matched on age and sex. For each subject, extensive questionnaire data are available, as well as one blood sample in which cotinine concentration has been measured by mass spectrometry–based methods.[30] These data provide detailed information on smoking history and a quantitative measurement of smoking intensity at enrollment. The main characteristics of the studied population are given in eTable 1 (http://links.lww.com/EDE/A742).

Exposure Assessment

As detailed in eAppendix (http://links.lww.com/EDE/A742) section 1, the questionnaire data describing smoking habits as a function of age were used to derive, for each participant at each year from birth to the end of follow-up, the average smoking intensity (measured in number of cigarettes smoked per day). We also accounted for a background exposure to tobacco smoke (mainly reflecting passive smoking) by sampling, for each participant at each year, a blood cotinine level from the cotinine distribution in nonsmokers at the time of blood collection (ie, never or former smokers). This concentration was subsequently converted in terms of fractional smoking intensity and added to the active smoking exposure (if any). The resulting exposure history consists of a cumulated smoking intensity for each individual i at each calendar year t: $E^i(t)$.

Parameterization of the Hidden Markov Model

Along the disease course, each study subject moves across four states: Susceptible (S), healthy persons; Incubating (I), persons with a growing and undiagnosed lesion/tumor; Removed (R), patients with a diagnosed lung cancer; and (M), persons who died from a cause other than lung cancer (Figure 1). We focus on the first diagnosis and consider state R as absorbing. By definition, states S and I are hidden and only their union ($S \cup I$) can be observed: symptom-free individuals can be either healthy or with an undiagnosed tumor.

Time (t) is considered as discrete in the model and the time unit is 1 year. The first time interval ($t = 1$) is defined as the year of birth of the oldest person in the study (1929), and the last interval is the year at which the last event (diagnosis or death from a cause other than lung cancer) was observed (2010).

The S to I transition occurs with the last irreversible event causing one cell's activity to be altered and ultimately to form a tumor. According to Knudson's hypothesis, this transition corresponds to the "last hit"[31,32] or to malignant conversion in the multistage model proposed by Moolgavkar and Luebeck.[33] This assumes that once the S to I transition occurs, target cells have been irreversibly affected by exposures and can multiply to eventually form a tumor only according to a dynamic that is not necessarily driven by the same factors (eg, the tumor growth process may not depend on exposures). The time spent in state I defines the time to diagnosis, which reflects both the dynamics of malignant cell multiplication through the time taken for the tumor to become detectable and the screening efficiency (ie, the time interval for someone with a detectable tumor to be tested and diagnosed). The Markovian property applied to compartmental models imposes that the time spent in each state is exponentially distributed. This parametric assumption may be too restrictive and may be relaxed by arbitrarily subdividing a given state into K substates.[26,34] Here, we

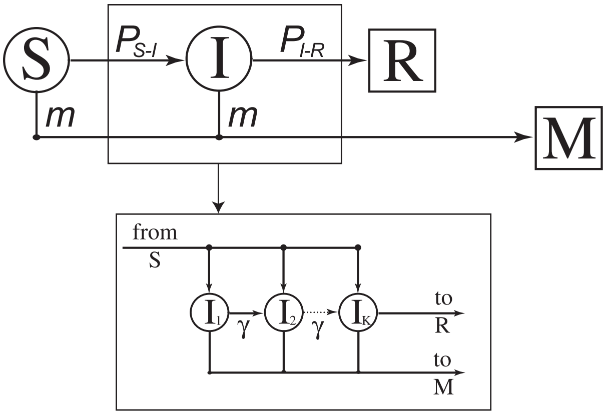

FIGURE 1. State space of the model. Circled states are hidden, others are observed.

consider K substates ($I_1, \ldots, I_K$) a person has to pass through to reach R. To enable any S–I_a, I_a–I_b, I_a–R, a, b € $[1, K]^2$ transition within a 1-year interval, time is considered as continuous in the subchain with transition rate γ. In the context of the present study, substates are included only for technical reasons (to ensure a flexible modeling of time to diagnosis), and the number of such substates K is fixed.

The corresponding 1-year interval transition can be expressed using a Gamma distribution with parameters γ and K. As a first approach, we considered disease progression to be independent of age. Sensitivity to the latter assumption has formally been assessed (eAppendix, http://links.lww.com/EDE/A742 section 4.2).

Other-cause mortality rates ($m^i(t)$ for individual i at time t) were derived from a publicly available actuarial table providing the mortality rates by age, sex, and smoking status (eAppendix, http://links.lww.com/EDE/A742 section 2.1).[35]

We define the probability of an S to I transition as:

$$p^i_{S-I}(t) = \frac{\exp\left[\mu + \lambda_1 a^i(t) + S^i(t)\lambda_2 a^i_0 + \lambda_3 t^i_q(t)\right]E^i(t)}{1+\exp\left[\mu + \lambda_1 a^i(t) + S^i(t)\lambda_2 a^i_0 + \lambda_3 t^i_q(t)\right]E^i(t)}\left(1 - m^i(t)\right),$$

where $S^i(t)$ is the binary smoking status for person i at time t, $t^i_q(t)$ is the time elapsed since smoking cessation, and μ, λ_1, λ_2, and λ_3 are four real parameters. The probability of entering lung carcinogenesis, $p^i_{S-I}(t)$, is then defined as a function of exposure to tobacco smoke $E^i(t)$ and can be decomposed into four terms: (1) an intercept on the logistic scale measured by μ; (2) the effect of age at t $a^i(t)$ measured by λ_1; (3) the effect of age when started smoking $a^i_0(t)$ measured by λ_2; and (4) the effect of time since quitting smoking $t^i_q(t)$ measured by λ_3. For never-smokers, $S^i(t) = 0$ for all t and the effect of age at starting smoking is set to 0. For current and never-smokers, the effect of time since smoking cessation is null $\left(t^i_q(t) = 0\right)$. This function was chosen so that it yields a null probability in nonexposed persons (ie, those who were never actively or passively exposed to tobacco smoke), it is an increasing function of exposure, and it tends to be 1.0 for an infinite exposure.

To address the issue of temporal collinearity between age and exposure duration, we consider lifetime cumulative exposure functions. Hence, effects of age, age at starting smoking, and time since smoking cessation (estimated through λ_1, λ_2, and λ_3, respectively) are adjusted for exposure duration.

The individual exposure history $E^i(t)$ is considered to be quasi-observed, plugged into the model, and used for the formal estimation of the five parameters μ, λ_1, λ_2, λ_3, and γ.

The definition of the full set of transition probabilities, together with details of the likelihood calculations, are given in the eAppendix (http://links.lww.com/EDE/A742), including our general recursive procedure to exactly calculate the longitudinal probability to be asymptomatically affected.[26] Parameter estimation was done through a Metropolis-Hastings algorithm detailed in eAppendix (http://links.lww.com/EDE/A742) section 3, setting uninformative uniform prior distributions on [–100, 100] for μ, λ_1, λ_2, λ_3, and $\log(\gamma)$.

Based on the joint posterior distribution of μ, λ_1, λ_2, λ_3, and γ, obtained from the Markov Chain Monte Carlo run, it is then possible to simulate individual trajectories across the unobserved states S, I, R, and M. Such simulations have the potential to provide estimates of the model fit because they quantify how well the model is reconstructing each of the individual trajectories and its dynamics: through simulated transitions and calendar year at which these occurred. To account for variability in the parameter estimates, we ran the simulation for 10,000 sets of parameters sampled from their joint posterior distribution. Simulations were summarized by the mean time spent in I (ie, time to diagnosis) and the proportion of simulations for which an S to I transition was simulated (p_{case}) in each participant, given the person's exposure and main risk determinants.

RESULTS

The Markov Chain Monte Carlo algorithm ran for 50,000 iterations, and the convergence of the runs was visually assessed from the history plots for each parameter.

Parameters Estimate—Model Assessment

Results setting $K = 2$ are summarized in Table 1, where the posterior mean and 95% credible interval are given for each parameter. Corresponding posterior distributions are plotted in eFigure 2 (http://links.lww.com/EDE/A742) and show a sharp shape.

TABLE 1. Parameter Estimates for $K = 2$, Setting Each of λ_1, λ_2, and λ_3 to 0

	Parameter Estimates: Posterior Mean (95% Credible Interval)					Bayesian Information Criterion Score
	μ	λ_1	λ_2	λ_3	γ	
Full model	1.52 (0.44 to 2.86)	0.03 (0.01 to 0.05)	–0.08 (–0.10 to –0.06)	–0.06 (–0.08 to –0.05)	2.45 (2.13 to 2.81)	8255.6
$\lambda_1 = 0$	3.32 (3.01 to 3.62)	—	–0.08 (–0.10 to –0.06)	–0.06 (–0.08 to –0.04)	2.56 (2.22 to 2.93)	8257.0
$\lambda_2 = 0$	1.86 (0.31 to 3.55)	0.03 (0.00 to 0.05)	—	–0.11 (–0.13 to –0.09)	3.27 (3.00 to 3.57)	8287.4
$\lambda_3 = 0$	2.22 (1.28 to 3.21)	0.01 (0.00 to 0.03)	–0.11 (–0.12 to –0.09)	—	2.29 (1.94 to 2.68)	8301.5

Results (posterior mean and 95% credible intervals) are based on 50,000 iterations (with 20,000 iterations burn-in).

μ indicates intercept on the log scale; λ_1, effect of age $a^i(t)$; λ_2, effect of age at starting smoking a^i_0; λ_3, effect of time since smoking cessation $t^i_q(t)$; and γ, continuous time I_a–I_b transition rate.

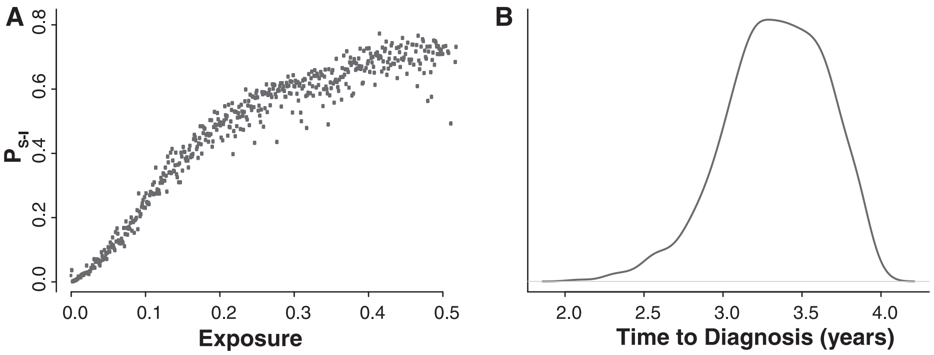

FIGURE 2. Summary of the 10,000 simulated trajectories from the joint posterior distribution of μ, λ_1, λ_2, λ_3, and γ estimated setting $K = 2$. A, Dose-response curves. For clarity, the exposure has been discretized, and whenever several observations were available in one range of exposure, the median value of p_{S-I} has been plotted. B, Density estimation of the time to diagnosis (posterior mean = 3.2 years).

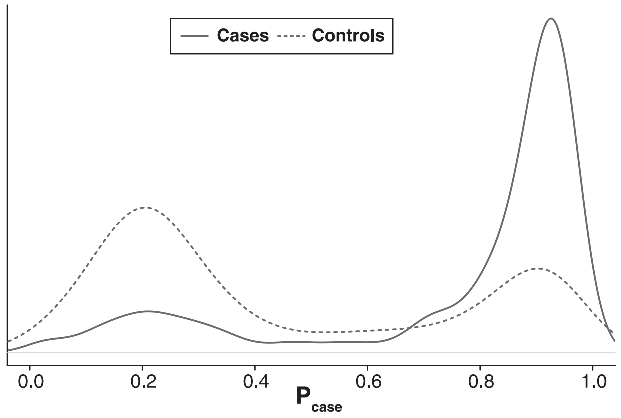

FIGURE 3. Density estimation of the probability of simulating an S to I (p_{case}) transition in cases (solid line) and controls (dashed lines). Results are presented for $K = 2$ and are based on 10,000 simulations derived from the joint posterior distribution of μ, λ_1, λ_2, λ_3, and γ.

Our model estimated a positive effect of age (measured by λ_1) and a negative effect of age at starting smoking (measured by λ_2) and of time since smoking cessation (measured by λ_3). The contributions of λ_1, λ_2, and λ_3 to the fit of the model were assessed by running the models with each of these parameters sequentially set to 0 and comparing the quality of the fit—measured by their Bayesian information criterion (BIC) scores—of resulting models to that of the full model. As shown in Table 1, the model with $\lambda_1 = 0$ yields BIC values that are very close to those obtained for the full model (differences in BIC lower than 2), suggesting that including λ_1 in the model only marginally improves the quality of fit. Conversely, setting λ_2 or λ_3 to 0 leads to a greater increase of the BIC compared with the full model (differences in BIC greater than 30 and 45, respectively). This suggests that, once adjusted on exposure duration, age has a weak independent effect on the risk of smoking-induced lung cancer. In contrast, age at starting smoking for ever-smokers and time since smoking cessation for former smokers have a stronger negative effect, leading early smokers and late quitters to be at greater risk of lung cancer irrespective of exposure levels.

Simulations of individual trajectories are summarized in Figure 2, which shows the median of p_{S-I} (over 10,000 simulated values), for each person/calendar year combination, as a function of the exposure estimated for that person in that year (Figure 2A). This plot suggests a leveling-off of the dose-response relationship and a saturation of the risk of lung cancer at high tobacco smoke exposures. The distribution of the time spent in I is reported in Figure 2B and suggests a time to diagnosis ranging from 1 to 4 years.

The quality of the fit of our model can be assessed by analyzing its ability to reconstruct individual trajectories among cases and controls separately. Figure 3 shows that the average probability of simulating an S to I transition (p_{case}) in actual cases peaks at 93% while a secondary mode can be observed around 20%. eFigure 3 (http://links.lww.com/EDE/A742) shows that the latter corresponds to never-smokers, whose trajectory is by default not well reconstructed by our model that considers tobacco smoke as the only risk factor for lung cancer. The distribution of p_{case} in controls is left-tailed, and its mode is around 21%. Among controls in whom an S to I transition was frequently simulated, a vast majority were heavy smokers; more than 95% of controls with $p_{\text{case}} \geq$ 25% were ever-smokers. These are typically high-risk and yet disease-free individuals. We also investigated the ability of our model to reproduce the dynamics of disease progression. Simulations showed satisfactory performances, with an average time gap between simulated and actual date of diagnosis of 2.3 years (95% credible interval = 1.3 to –3.3).

Sensitivity Analyses

The model was run for three other values of K ($K = 5$, 10, and 15), and resulting parameter estimates are summarized in Table 2. Estimates of λ_1, λ_2, and λ_3 seem unaffected by the choice of K, whereas estimates of γ clearly decrease with the number of hidden states. As expected, when K is larger, the number of required I_a–I_b transitions to reach R increases, in turn constraining γ so that the overall time spent in I is consistent across all values of K examined. We found that the time taken for a tumor to be detected ranged from 1.0 to 4.0 years, with limited overlap across simulations (eFigure 4, http://links.lww.com/EDE/A742). Longer times to diagnosis imply more asymptomatic cases in the population at censorship, hence supporting a larger number of S to I transitions and larger values of μ. This explains why the ranking of models

TABLE 2. Parameter Estimates for K = 2, 5, 10, and 15

	Parameter Estimates: Posterior Mean (95% Credible Interval)				
	μ	λ_1	λ_2	λ_3	γ
K = 2	1.52 (0.44 to 2.86)	0.03 (0.01 to 0.05)	–0.08 (–0.10 to –0.06)	–0.06 (–0.08 to –0.05)	2.45 (2.13 to 2.81)
K = 5	0.63 (–0.27 to 1.33)	0.03 (0.01 to 0.04)	–0.11 (–0.13 to –0.09)	–0.03 (–0.04 to –0.02)	0.02 (0.00 to 0.13)
K = 10	0.97 (0.25 to 1.75)	0.02 (0.01 to 0.03)	–0.11 (–0.12 to –0.09)	–0.03 (–0.05 to –0.02)	0.01 (0.00 to 0.05)
K = 15	1.24 (0.42 to 2.03)	0.02 (0.01 to 0.03)	–0.11 (–0.12 to –0.09)	–0.04 (–0.05 to –0.03)	0.01 (0.00 to 0.04)

Results (posterior mean and 95% credible intervals) are based on 50,000 iterations (with 20,000 iterations burn-in).

μ indicates intercept on the log scale; λ_1, effect of age $a^i(t)$; λ_2, effect of age at starting smoking a_0^i; λ_3, effect of time since smoking cessation $t_q^i(t)$; and γ, continuous time I_a–I_b transition rate.

with respect to estimates of μ is consistent with the one based on time to diagnosis.

Models for K = 2, 5, 10, and 15 were compared on the basis of their ability to reconstruct trajectories in both cases and controls. There is a natural trade-off between simulating an S to I transition in actual cases and not simulating such a transition in controls. eFigure 5 (http://links.lww.com/EDE/A742) clearly shows that setting K = 2 yields better sensitivity in cases at the cost of slightly lower performances in controls. Altogether this suggests that setting K = 2 provides a better balance between these two antagonistic features.

Robustness of our results to the prior specification was assessed by substituting the uniform prior (with support [–100;100]) with zero-centered Gaussian priors for μ, λ_1, λ_2, λ_3, and log(γ). As summarized in eTable 2 (http://links.lww.com/EDE/A742), we considered prior variances ranging from 1,000 to 10. Results clearly show that none of the parameter estimates (except for λ_1, very marginally) was affected by the prior choice.

We also generalized our model to (1) enable a more flexible modeling of the role of exposure in $p_{S\text{–}I}$ (through an additional parameter λ_0), and (2) account for an age-related time to diagnosis (through an additional parameter θ), as detailed in eAppendix (http://links.lww.com/EDE/A742) sections 4.1 and 4.2. For $\lambda_0 = 1$, the S to I transition probability corresponds exactly to the reference model described above in Equation 1. For $\theta = 0$, the time spent in I is considered independent of age, which again corresponds to the model described above. The comparisons of parameter estimates for the fully generalized model (ie, including θ and λ_0) and the model in which $\theta = 0$ on the one hand, and for the model in which $\lambda_0 = 1$ and the model in which $\lambda_0 = 1$ and $\theta = 0$, on the other hand, both show that estimates of μ and γ are affected only by the inclusion of an age-dependent process driving the transitions among the substate of I (eTable 3, http://links.lww.com/EDE/A742). Both models including θ show a moderate and positive effect of age, resulting in older individuals being diagnosed earlier after tumor initiation. However, corresponding BIC scores suggest marginal improvement of the fit (BIC differences <9). Consistently, simulations show that including an age-dependent sojourn time in the hidden state I does not yield any substantial improvement in the model ability to reconstruct trajectories in either cases or controls (eFigure 6, http://links.lww.com/EDE/A742).

Simulations also showed (eFigure 6, http://links.lww.com/EDE/A742) that the baseline model (in which $\theta = 0$ and $\lambda_0 = 1$) performed better in reconstructing trajectories of cases than the model including the effect of exposure ($\lambda_0 \neq 1$), at the cost of slightly lower specificity in controls, exemplified by a wider peak at higher values of p_{case}. Given that in our study population controls are twice as numerous as cases, and that each control contributes more to the likelihood than cases, the better performances of the model including λ_0 in reconstructing trajectories in controls yield an overall better fit to the data as measured by a lower BIC score compared with the baseline model (8,174.0 and 8,247.5, respectively).

Our data included age at recruitment as a case-control matching criterion, resulting in the age distribution in controls being right-shifted compared with that of the full cohort population (eFigure 7, http://links.lww.com/EDE/A742). As detailed in eAppendix (http://links.lww.com/EDE/A742) section 4.3, we performed a sensitivity analysis by resampling subsets of controls with and without age matching. Results showed, as expected, estimates of λ_1 based on unmatched data to be consistently higher than those based on the unmatched data (eTable 4, http://links.lww.com/EDE/A742). Nevertheless, in both scenarios, the inclusion of attained age yielded only moderate improvements in the model fit (differences in BIC <10), suggesting that our conclusion regarding the absence of an effect of age (other than through exposure duration) on the probability of entering carcinogenesis was not affected or driven by age matching.

DISCUSSION

We have developed an individual-based compartmental model to estimate parameters driving the dynamics of lung cancer progression. Our model for the probability to enter carcinogenesis accounted for the direct effect of age on exposure by considering lifetime cumulative exposure functions. We defined a logistic model for this probability, and subsequent simulations showed accurate trajectory reconstruction in cases and in the vast majority of controls. This model was generalized and included a more flexible modeling of the role of exposure on the probability of entering carcinogenesis. Although this model provided a better fit to the data, it did not yield better performances in reconstructing individual trajectories.

By construction, our model did not perform well in predicting disease onset in cases with low exposed cases. To improve the model, causes other than smoking should be accounted for by including main exposures (eg, occupational and environmental) and risk factors (eg, genetic polymorphisms). Such refinements would be enabled by the structural flexibility of our model, which can accommodate both time-dependent (eg, exposure history) and constant risk determinants (eg, disease risk genetic markers or a single measurement of other "-omic" markers) in the S to I transition probability.

We show that, once the direct effect of age on exposure duration has been accounted for, age itself has no further impact on the probability of initiating lung carcinogenesis, whereas age at starting smoking appears to be an influential covariate. In accordance with the exposome paradigm[2] and the idea that exposures at critical life stages have differential effects, this result highlights age at first active exposure to tobacco smoke as a driver for the risk of lung cancer. Higher lung cancer relative risks have been commonly found to be associated with earlier ages at initiation.[36,37] After several recent studies based on a two-stage clonal expansion model,[24,38–40] we considered smoking duration as a driver for the risk of initiating lung carcinogenesis. Despite modeling its effect directly, we considered lifetime cumulative exposure estimates that were subsequently plugged into the model as quasi-observations. Our approach then incorporates the effect of age at starting smoking and time since smoking cessation on exposure duration. Based on individual trajectories, our model provides estimates of the adjusted effect of age at starting smoking on the risk of lung cancer that support the existence of susceptibility to tobacco smoke based on early exposure. Similarly, we found the time since quitting smoking to have a protective effect on the risk of lung cancer, as previously reported in the literature.[41]

The probability of developing a tumor in highly exposed persons was estimated to plateau, which is consistent with the leveling-off of the relative risk demonstrated in previous studies in heavy smokers.[42] Possible reasons include potential saturation effects, dose-dependent inhalation habits, and possibly depletion of susceptible subjects or increased measurement error at high exposures. Although our finding is also consistent with a prominent and saturating effect of smoking duration in malignant cell promotion, our model, in its current form, is not able to identify which step of the carcinogenic process is mostly affected. However, the described model theoretically could be extended to incorporate transitions between states representing cellular physiologic changes involved in cancer development.

Our study also provides insights into the dynamics of lung cancer pathogenesis and shows that patients are diagnosed 1 to 4 years after putative malignant conversion. This is consistent with the lag-time reported between malignant conversion and death from lung cancer among miners exposed to arsenic, radon, or cigarette and pipe smoke from a biologically based two-stage clonal expansion model.[43]

In additional sensitivity analyses, our results were robust to the assumption that the time spent in I was independent of age and therefore suggested that our data did not support an age-dependent time to diagnosis. In the current setting, the role of clinical screening and technical detection efficiency as drivers of the time to diagnosis cannot be ruled out because substates I_1, …, I_K do not have a biological meaning. However, this could be modeled by considering that the first j substates correspond to the time needed for the tumor to become detectable and that the remaining K–j states relate to the time needed for a patient with a detectable tumor to be screened and diagnosed.

We could also relax the model from the assumption that all participants are in state S at enrollment by including the initial state among S, I_1, …, I_K in the sampling scheme. Biomarkers of disease onset (eg, genes whose expression is different in diagnosed cases) could be used to inform the distribution of the health status at enrollment.

We showed that the application of a compartmental model to reconstruct the course of smoking-induced lung cancer provides biologically valid results and enables the investigation of multiple and dynamic aspects of disease risk. Although latest developments of regression-based models are also able to integrate the full individual exposure history in risk estimation[44] and provide well-established measures of association, our approach shows complementary advantages with respect to modeling and parametric flexibility and refined measures of the performances of the model. The longitudinal nature of our model also allows age-dependent susceptibility functions to be included as disease risk determinants, whose estimation would constitute an intuitive approach for the identification of critical life stages at which each exposure is driving the risk of disease onset.

Based on these properties, we believe that the present study shows the potential for the application of longitudinal models for the life-course risk of chronic diseases.

ACKNOWLEDGMENTS

We thank Martyn Smith, Lutzen Portengen, and Benoît Liquet for their very useful and constructive inputs on the model.

REFERENCES

1. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005;14:1847–1850.
2. Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. *Science*. 2010;330:460–461.

3. Vermeulen R, Chadeau-Hyam M. Dynamic aspects of exposure history-do they matter? *Epidemiology*. 2012;23:900–901.
4. Ben-Shlomo Y, Kuh D. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *Int J Epidemiol*. 2002;31:285–293.
5. Lubin JH, Alavanja MC, Caporaso N, et al. Cigarette smoking and cancer risk: modeling total exposure and intensity. *Am J Epidemiol*. 2007;166:479–489.
6. Lubin JH, Caporaso N, Wichmann HE, Schaffrath-Rosario A, Alavanja MC. Cigarette smoking and lung cancer: modeling effect modification of total exposure and intensity. *Epidemiology*. 2007;18:639–648.
7. Galea S, Riddle M, Kaplan GA. Causal thinking and complex system approaches in epidemiology. *Int J Epidemiol*. 2010;39:97–106.
8. Commenges D. Multi-state models in epidemiology. *Lifetime Data Anal*. 1999;5:315–327.
9. Frydman H. A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to aids. *J Roy Stat Soc*. 1992;54:853–866.
10. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Stat Med*. 1997;16:2191–2210.
11. Guihenneuc-Jouyaux C, Richardson S, Longini IM Jr. Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline. *Biometrics*. 2000;56:733–741.
12. Longini IM Jr, Clark WS, Gardner LI, Brundage JF. The dynamics of CD4+ T-lymphocyte decline in HIV-infected individuals: a Markov modeling approach. *J Acquir Immune Defic Syndr*. 1991;4:1141–1147.
13. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
14. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol*. 2002;155:176–184.
15. Joffe M, Gambhir M, Chadeau-Hyam M, Vineis P. Causal diagrams in systems epidemiology. *Emerg Themes Epidemiol*. 2012;9:1.
16. Moolgavkar SH, Venzon DJ. Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Math Biosci*. 1979;47: 55–77.
17. Moolgavkar SH, Knudson AG Jr. Mutation and cancer: a model for human carcinogenesis. *J Natl Cancer Inst*. 1981;66:1037–1052.
18. Loeb LA, Loeb KR, Anderson JP. Multiple mutations and cancer. *Proc Natl Acad Sci U S A*. 2003;100:776–781.
19. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc Natl Acad Sci U S A*. 2002;99: 15095–15100.
20. Meza R, Jeon J, Moolgavkar SH, Luebeck EG. Age-specific incidence of cancer: phases, transitions, and biological implications. *Proc Natl Acad Sci U S A*. 2008;105:16284–16289.
21. Schöllnberger H, Beerenwinkel N, Hoogenveen R, Vineis P. Cell selection as driving force in lung and colon carcinogenesis. *Cancer Res*. 2010;70:6797–6803.
22. Deng L, Kimmel M, Foy M, Spitz M, Wei Q, Gorlova O. Estimation of the effects of smoking and DNA repair capacity on coefficients of a carcinogenesis model for lung cancer. *Int J Cancer*. 2009;124:2152–2158.
23. Foy M, Spitz MR, Kimmel M, Gorlova OY. A smoking-based carcinogenesis model for lung cancer risk prediction. *Int J Cancer*. 2011;129: 1907–1913.
24. Heidenreich WF, Wellmann J, Jacob P, Wichmann HE. Mechanistic modelling in large case-control studies of lung cancer risk from smoking. *Stat Med*. 2002;21:3055–3070.
25. Kaiser JC, Heidenreich WF. Comparing regression methods for the two-stage clonal expansion model of carcinogenesis. *Stat Med*. 2004;23:3333–3350.
26. Chadeau-Hyam M, Clarke PS, Guihenneuc-Jouyaux C, Cousens SN, Will RG, Ghani AC. An application of hidden Markov models to the French variant Creutzfeldt–Jakob disease epidemic. *J Roy Stat Soc*. 2010;59:839–853.
27. Palli D, Berrino F, Vineis P, et al; EPIC-Italy. A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *Tumori*. 2003;89:586–593.
28. Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann Oncol*. 1992;3:783–791.
29. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;5(6B):1113–1124.
30. Johansson M, Relton C, Ueland PM, et al. Serum B vitamin levels and risk of lung cancer. *JAMA*. 2010;303:2377–2385.
31. Knudson AG. Two genetic hits (more or less) to cancer. *Nat Rev Cancer*. 2001;1:157–162.
32. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*. 1971;68:820–823.
33. Moolgavkar SH, Luebeck G. Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. *Risk Anal*. 1990;10:323–341.
34. Garske T, Ward HJ, Clarke P, Will RG, Ghani AC. Factors determining the potential for onward transmission of variant Creutzfeldt-Jakob disease via surgical instruments. *J R Soc Interface*. 2006;3:757–766.
35. SOA. Experience Studies-Individual Life—2008 Valuation Basic Tables [VBT] Report and Tables. Available at: http://www.soa.org/research/experience-study/ind-life/valuation/2008-vbt-report-tables.aspx. Accessed 1 January 2008.
36. Hoggart C, Brennan P, Tjonneland A, et al. A risk model for lung cancer incidence. *Cancer Prev Res (Phila)*. 2012;5:834–846.
37. Pesch B, Kendzia B, Gustavsson P, et al. Cigarette smoking and lung cancer—relative risk estimates for the major histological types from a pooled analysis of case-control studies. *Int J Cancer*. 2012;131:1210–1219.
38. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ*. 2000;321:323–329.
39. Vineis P, Kogevinas M, Simonato L, Brennan P, Boffetta P. Levelling-off of the risk of lung and bladder cancer in heavy smokers: an analysis based on multicentric case-control studies and a metabolic interpretation. *Mutat Res*. 2000;463:103–110.
40. Hazelton WD, Luebeck EG, Heidenreich WF, Moolgavkar SH. Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model. *Radiat Res*. 2001;156:78–94.
41. Richardson DB, Cole SR, Langholz B. Regression models for the effects of exposure rate and cumulative exposure. *Epidemiology*. 2012;23: 892–899.
42. Vineis P, Kogevinas M, Simonato L, Brennan P, Boffetta P. Levelling-off of the risk of lung and bladder cancer in heavy smokers: an analysis based on multicentric case-control studies and a metabolic interpretation. *Mutat Res*. 2000;463:103–110.
43. Hazelton WD, Luebeck EG, Heidenreich WF, Moolgavkar SH. Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model. *Radiat Res*. 2001;156:78–94.
44. Richardson DB, Cole SR, Langholz B. Regression models for the effects of exposure rate and cumulative exposure. *Epidemiology*. 2012;23:892–899.