Supporting Information - Longitudinal Analysis of the Risk of Smoking-Induced Lung Cancer: A Compartmental Hidden Markov Model

The `C++` code developed for this study (`QOMiC`) and its documentation are freely available at the following address

http://imperial.ac.uk/people/m.chadeau/

1 Exposure Assessment

In this section we describe how exposure is derived from both questionnaire data and the single cotinine concentration measurement available for each participant at time of enrollment. We define one cumulative (over lifetime) exposure function expressed in terms of yearly averaged smoking intensity (in number of cigarettes smoked per day). These estimates are subsequently plugged into the model and considered as fixed.

1.1 Reconstructing individual yearly smoking history

Data from the lung cancer case-control study nested within EPIC consists in 757 cases and 1524 controls matched on age and gender who were enrolled before clinical onset. Specifics of the studied population are summarized in Table S1. For each participant, questionnaire-based data describing individual smoking history is available. Specifically:

- the smoking status (never, former, or current smoker)
- the age at starting smoking, for ever smokers
- the age at quitting smoking, for former smokers
- the detailed description of active smoking episodes (for non-continuous smokers), including the age at starting and stopping smoking for each smoking period.
- the smoking intensity (average daily smoking intensity) at enrollment
- the smoking intensity per decade of age (for ever smokers).

Assuming that individual smoking habits have remained unchanged since the last available follow-up, we derived from these data the detailed yearly smoking history: we first identified the calendar years of active smoking (if any), as well as the corresponding average smoking intensity. To account for both misreporting and yearly variation in smoking habits within a given decade of age, we arbitrarily introduced a year-by-year variability in the smoking intensity. Denoting $r^i(t)$ the reported smoking intensity for individual i at calendar year t, we sampled the actual smoking intensity $s^i(t)$ from a Gaussian distribution centered on $r^i(t)$, and with a variance defined such that the width of the (Gaussian) 95^{th} confidence interval is 2 cigarettes/day if $r^i(t) \leq 10$, 5 cigarettes/day if $10 < r^i(t) \leq 35$, and 10 cigarettes/day if $r^i(t) > 35$.

1.2 Estimating the smoking intensity *vs.* cotinine concentration relationship

One snapshot measurement of cotinine concentration in blood is also available for each participant. The blood sample from which the cotinine titre has been measured was prospectively collected. In Fig. S1-a, we plot the cotinine concentration as a function of the reported smoking intensity in those who smoked at time of blood collection (N=736). This plot clearly shows a levelling-off of the intensity-to-cotinine relationship, that could be due to a saturation of blood cotinine and/or to under-reporting in heavy smokers. To model these data, we fitted the following linear model linking the mean cotinine levels per class of number of cigarettes smoked and the logarithm of the reported smoking intensity r:

$$\alpha + \beta \log(r) \text{ if } \; r \geq 1 \text{ cig/day} \tag{1}$$

The model was fitted on observations from current smokers only. Resulting estimates $\hat{\alpha}$=72.71 and $\hat{\beta}$=431.97 provided a good fit to the data, with $R^2 > 85\%$.

1.3 Deriving the individual exposure function

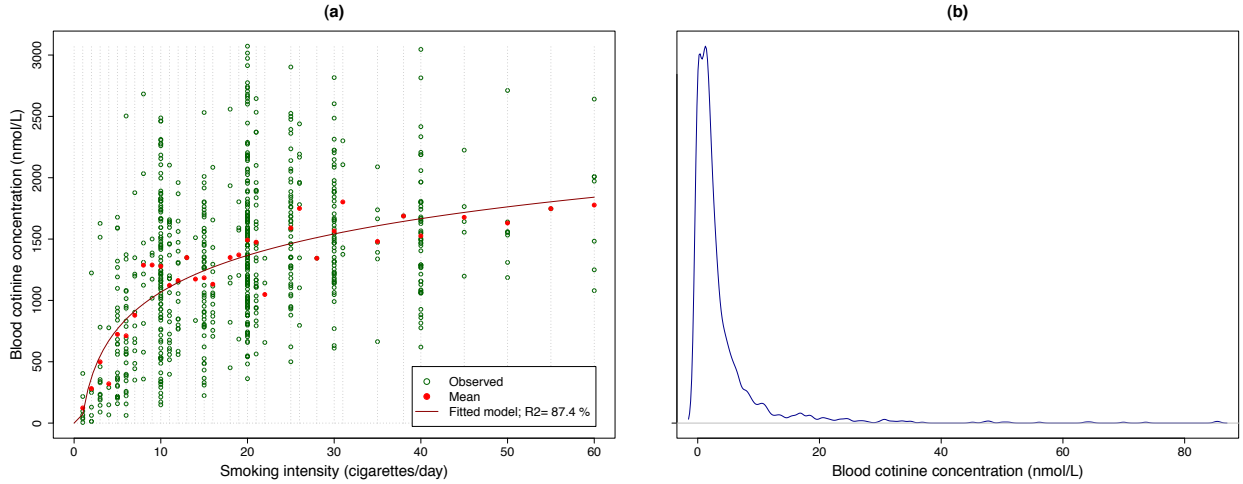
The exposure function is obtained from the set of values $s^i(t)$, which represent, for each individual at each calendar year, the yearly averaged smoking intensity. We accounted for a 'background exposure' reflecting, for instance, passive smoking and sampled, for every individuals at every calendar year, the background cotinine exposure ($b^i_{cot}(t)$) from the empirical distribution of the cotinine level in non-smokers at time of blood collection (Fig. S1-b). This background cotinine exposure was then translated into a fractional number of cigarettes smoked per day, assuming a linear relationship between cotinine and smoking intensity in non-smokers ($r < 1$ cig/day).

The cumulative cigarette exposure for individual i at time t is then defined as:

$$E^i(t) = \sum_{u=t^i_0}^{t} \left(s^i(u) + b^i_{cot}(u)/\hat{\alpha} \right), \tag{2}$$

where t^i_o is the year of birth of individual i.

Figure S1: Fig. S1-a: Blood cotinine concentration (nmol/L) *vs.* reported smoking intensity (#cigarettes/day) for smokers at time of inclusion (N=736). The fitted curve is based on the mean cotinine concentration per smoking intensity classes. Fig. S1-b: Density estimation of the blood cotinine concentration in non-smokers at the time of blood collection (*i.e.* never or former smokers; N=1545).



2 Model Description – Likelihood Calculation

2.1 Transition Probabilities

In our model we consider four states:

- S: healthy individuals;
- I: 'incubating' individual (with a growing and undiagnosed tumor);
- R: clinical individuals (whose tumor has been diagnosed);
- M: deceased individuals (from a cause other than lung cancer).

In this analysis we focus on the time to first diagnosis, and hence consider R as an absorbing state. We do not include death from lung cancer in the state space, which would require the inclusion of a remission state and to allow for backward transitions from remission to S. This would also necessitate the modeling of both treatment effect and survival, which is theoretically feasible, but would require information on the type of treatment and on the efficiency of different treatments.

In our simpler setting, no backward transitions are considered, and the only non-zero transition probabilities are listed and defined below.

The transition from S to M represents death from another cause for an individual with no lung tumor. This probability is fixed in the model and has been derived from publicly available actuarial data giving the instant mortality in the general insured population by age, gender and smoking status (VBT tables). Assuming that the US insured population is comparable to the European population included in EPIC, we computed for each individual i at each year t, the mortality rate $m^i(t)$ conditional on survival until $t-1$. We make the assumption that having entered carcinogenesis without being diagnosed does not impact the other-cause mortality and set the I to M transition probability to $m^i(t)$ for individual i at time t. This simplifying assumption can be justified by the fact that one undiagnosed lung cancer case (in I) would typically contribute to the estimation of the mortality rate in the original table.

The S-I transition happens with the last irreversible event causing the metabolism of one cell to be modified, leading to a malignant cell and ultimately to a tumor. For individual i at time t, this probability is defined as:

$$p^i_{S-I}(t) = \frac{\exp\left(\mu + \lambda_1 a^i(t) + S^i(t)\lambda_2 a^i_0 + \lambda_3 t^i_q(t)\right) E^i(t)}{1 + \exp\left(\mu + \lambda_1 a^i(t) + S^i(t)\lambda_2 a^i_0 + \lambda_3 t^i_q(t)\right) E^i(t)}(1 - m^i(t)), \tag{3}$$

where μ is a parameter measuring the intercept of the model on the logistic scale, $a^i(t)$ is the age of individual i at time t, a^i_0 is the age at which individual i started smoking, $S^i(t)$ indicates the (binary) smoking status of individual i at time t, $t^i_q(t)$ is the time since smoking cessation for individual i at time t, and $m^i(t)$ is the other cause mortality for individual i at time t. We opted for this simple model because (i) it yields a null probability for non-exposed individuals; (ii) it is an increasing function of exposure; and (iii) it tends to 1 for an infinite exposure. This function is flexible enough to fit linear or curved dose-response relationships. The probability to enter lung carcinogenesis is upper bounded by $(1-m^i(t))$, in order to ensure stochasticity of the transition matrix.

The time spent in state I defines the time to diagnosis. To relax the model from the Markovian property according to which the time spent in a given state is exponentially distributed, and enable a flexible modelling of the time to diagnosis, we decompose state I into K sub-stages. Individuals must pass through each $I_k, k = 1, \ldots, K$ to reach R. The number of sub-states (K) is fixed and these sub-states have no biological interpretation. We allow any $I_a - I_b$, transition ($b \geq a$) within a one year interval by considering continuous time in the sub-chain and define the transition rate γ, which is constant over time and the same for any single jump transitions. Integrating over a one-year period, the yearly transition probability for individual i at time t is:

$$p^i_{I_a - I_b}(t) = \mathcal{P}(b-a, \gamma)(1 - m^i(t)), K \geq b > a \tag{4}$$

where

$$\mathcal{P}(b-a,\gamma) = \int_{u=0}^{1} \frac{1}{\Gamma(b-a)\gamma^{(b-a)}} u^{(b-a-1)} e^{-u/\gamma} du$$

is the cumulative (between 0 and 1) density function for the Gamma distribution with parameters $b-a$ and γ. Similarly, the probability to reach R is:

$$p^i_{I_a-R} = \mathcal{P}(K+1-a,\gamma)(1-m^i(t)), a \leq K. \tag{5}$$

The non-transition probabilities I_a-I_a, and S-S are defined as:

$$p^i_{I_a-I_a}(t) = 1 - \sum_{b\neq a} p^i_{I_a-I_b}(t)(1-m^i(t)), \tag{6}$$

and

$$p^i_{S-S}(t) = 1 - (p^i_{S-I}(t) + m^i(t)), \tag{7}$$

respectively.

2.2 Likelihood of the Hidden Markov Model

In our setting, states S and I are hidden, and only their union $SI = S \cup I$ can be observed. The contribution to the likelihood of individual i at time t can therefore be expressed as:

$$L^i(t) \propto \left\{ \left[1 - p^i_{SI-R}(t) - m^i(t)\right]^{n^i_{SI-SI}(t)} p^i_{SI-R}(t)^{n^i_{SI-R}(t)} \right\}, \tag{8}$$

where $n^i_{SI-SI}(t)$ and $n^i_{SI-R}(t)$ are two binary indicators such that $n^i_{SI-SI}(t)$=1 if individual i remains symptom-free at time t, and $n^i_{SI-R}(t)$=1 if individual i is diagnosed with lung cancer at time t. The only probability left in the kernel of the likelihood is $p^i_{SI-R}(t)$, which is to be expressed as a function of the modelled probabilities $p^i_{S-I}(t)$ and $p^i_{I-R}(t)$.

Denoting $H^i_k(t)$ the probability that individual i is in state I_k at time t, it comes:

$$p^i_{SI-R}(t) = \sum_{k=1}^{K} H^i_k(t-1) p^i_{I_k-R}(t), \tag{9}$$

where $p^i_{I_k-R}(t)$ is the probability that individual i makes the transition from I_k to R within the one-year interval t (Eq. [5]), and the function H_k is defined as follows.

2.3 Overview of the recursive calculation of the likelihood

Generalising the above notation and setting $H^i_S(t)$ as the probability that individual i is in state S at time t, $H^i_k(t+1)$ can be written as a weighted sum of transition probabilities from S to I_k, and from I_u to I_k:

$$H^i_k(t+1) = H^i_S(t) p^i_{S-I_k}(t) + \sum_{u=1}^{k} H^i_u(t) p^i_{I_u-I_k}(t). \tag{10}$$

The calculation of $H^i_S(t)$ is straightforward; an immediate recursion gives:

$$H^i_S(t) = \prod_{u=t^i_0}^{t-1} p^i_{s-s}(u), t > t^i_0, \tag{11}$$

where t^i_0 is the year at which individual i has been recruited. As a simplifying assumption, we consider that individuals are in S at their year of enrollment: $H^i_S(t^i_0) = 1$. Furthermore, we assume that once they move to I they enter the sub-chain in I_1: $p^i_{S-I_k}(t) = 0$, and $p^i_{S-I_1}(t) = p^i_{S-I}(t)$. Based on Eq. [10], we calculate $H^i_k(t)$ using a recursive procedure (over t and k) which is detailed below for a generic individual i:

1. Set $H_S^i(t_0^i) = 1$, $H_k^i(t_0^i)$=0, and t=$t_0^i + 1$
2. Calculate $H_S^i(t+j), j > 1$ from Eq. [11]
3. Set $k = 1$, and apply $H_1^i(t+1) = p_{S-I}^i(t^i)$
4. Increment t and apply Eq. [10];

$$H_1^i(t+1) = H_1^i(t) + H_S^i(t)p_{S-I}^i(t).$$

   Loop over t until t=t_f^i, the last year of follow-up for individual i.

5. Set $k = 2$, $t = t_0^i + 1$, and $H_k^i(t_0^i + 1) = 0$
6. Increment t and apply Eq. [10]:

$$\sum_{u=1}^{k} H_u^i(t)p_{I_u-I_k}^i(t)$$

   Loop over t until t=t_f^i, the last year of follow-up for individual i.

7. Increment k and repeat steps 5-6 until $k = K$
8. Derive $p_{SI-R}^i(t)$ from the result of step 7 using Eq. [9].
9. Calculate the set of contributions to the likelihood for each individual i, $L^i(t)$ using Eq. [8]
10. Repeat steps 1-9 for each included individual
11. Calculate the full likelihood:

$$L = \prod_{i,t} L^i(t)$$

3 Parametrisation of the MCMC Algorithm

We developed an MCMC algorithm based on the likelihood for the Hidden Markov model. A Metropolis-Hastings algorithm is used because it does not require the specification of the full conditional distribution of parameters. At each iteration, parameters are updated in the same order: μ, λ_1, λ_2, λ_3, and finally γ.

Candidate points for $\theta = (\mu, \lambda_1, \lambda_2, \lambda_3)$, denoted Θ, are sampled from a random walk proposal. At iteration i

$$\Theta^{(i)} \hookrightarrow \mathcal{N}(\theta^{(i-1)}, \sigma_\theta^2)$$

where σ_θ is fixed and $\theta^{(i-1)}$ is the previous draw from the chain.

To ensure that candidates for γ denoted Γ, are positive, we use a random walk for $\log(\gamma)$:

$$\log\left(\Gamma^{(i)}\right) \hookrightarrow \mathcal{N}(\log(\gamma^{(i-1)}), \sigma_\gamma^2),$$

where σ_γ is fixed, and $\gamma^{(i-1)}$ is the retained value at iteration $i-1$.

We assume uniform prior distributions for θ and γ. Combined with the symmetry of the normal proposal distributions, candidate points $\Theta^{(i)}$ are accepted with probability defined as the ratio of likelihood computed using the new candidate over the likelihood at the previous step of the algorithm. Candidates for $\Gamma^{(i)}$ are accepted with probability

$$r_\gamma = \min\left(1, \frac{L(\Theta^{(i)}, \Gamma^{(i)})}{L(\Theta^{(i)}, \gamma^{(i-1)})} \frac{\Gamma^{(i)}}{\gamma^{(i-1)}}\right).$$

To ensure that candidates are also sampled from the tail of the posterior distribution, σ_θ and σ_γ are set such that the acceptance rates of all parameters lie around 30%.

4 Sensitivity analyses

4.1 A more flexible model for p_{S-I}

We considered a more flexible model for the probability to enter carcinogenesis:

$$p^i_{S-I}(t) = \frac{\exp\left(\mu + \lambda_0 \log(E^i(t)) + \lambda_1 a^i(t) + S^i(t)\lambda_2 a^i_0 + \lambda_3 t^i_q(t)\right)}{1 + \exp\left(\mu + \lambda_0 \log(E^i(t)) + \lambda_1 a^i(t) + S^i(t)\lambda_2 a^i_0 + \lambda_3 t^i_q(t)\right)}(1 - m^i(t)), \tag{12}$$

where λ_0 is an additional parameter modelling the effect of exposure. This model shares the same parametrization as the one described in Eq[3]. More specifically, constraining λ_0=1 in Eq. (12) corresponds exactly to the model described in Eq. (3). Estimation of the additional parameter λ_0 is done via the MCMC procedure already described. The proposal distribution is a random walk and the prior distribution is set to an Uniform distribution with support [-100;100].

4.2 Modelling age-dependent sojourn time in I

For simplicity, γ, the transition rate driving any of the I_i-I_j and I_K-R transitions was considered independent of age. In order to assess how sensitive our results are to this assumption, we generalised our model such that the actual transition rate depends on age, assuming a linear relationship on the log scale. This enables to consider an age-dependent sojourn time in the hidden state I, in turn corresponding to the tumor growth process being different at different ages, and/or an age-dependent efficiency in screening and detection. Specifically we defined

$$\log(\gamma(a^i(t))) = log(\gamma_0) + \theta a^i(t),$$

where $a^i(t)$ is the age of individual i at time t, $\gamma^i(a^i(t))$ is the actual transition rate used to calculate transition probabilities $p^i_{I_i-I_j}(t)$, $p^i_{I_i-R}$, $p^i_{I_i-I_j}(t)$, $(i,j \in [1,K]; j \leq i)$ (see Eqs (4), (5), (6)). The background rate γ_0 corresponds to γ in the original model, and θ is an additional parameter to be estimated which defines the strength of the relationship. Estimation of θ is done by adding one step to the Metropolis-Hastings procedure, and as for the other real parameters, its sampling scheme involves a random walk in which the variance of the proposal is tuned to ensure an acceptance rate lying between 20 and 30%. If θ is set to 0, $\gamma(a^i(t)) = \gamma_0$, which corresponds to the age-independent time to diagnosis assumption embedded in the reference model.

Results (Table S3 and Fig S6) suggest that, irrespective of the exposure model considered, incorporating an age-dependent time to diagnosis only slightly improves the quality of fit. Consistently, simulations based on the joint posterior distribution of the parameters did not show any improvement in the predictive performances of the model. While this generalisation yields additional modelling flexibility, it requires constant updating of the transition matrix between sub-states of I introducing a substantial computational burden: in practice the algorithm was slowed one order of magnitude down.

Altogether this suggest that incorporating this refinement in our model will only results in a slight improvement of the quality of fit at the cost of considerable computational effort.

4.3 Assessing the impact of age matching on the estimates of λ_1

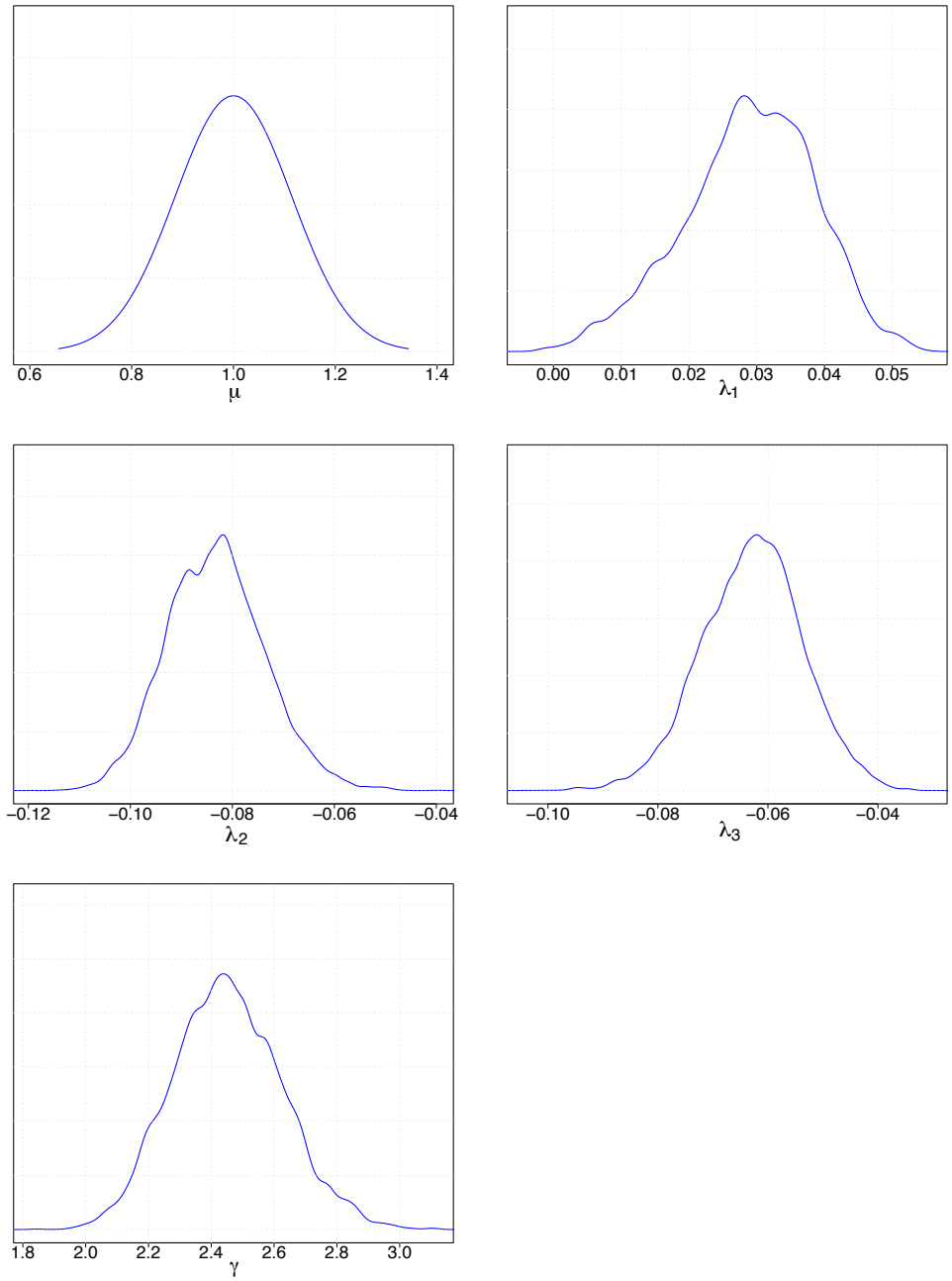
Case-control data retained in our analyses comprise (N=738) lung cancer cases and (N=1524) controls which where matched on age at recruitment. This matching mainly results in controls having an age at recruitment distribution different than the one from the full EPIC population (SI Fig S7). To measure the effect of that right-shifted age distribution in controls, we re-sampled (N=738) controls from our study population under two scenarios: (i)- imposing the same age matching as in the original data set, and (ii)- ensuring that the age distribution of controls was similar to that of the entire EPIC population. Twenty re-samples were independently drawn for both scenarios, and in each sampled population, we ran the model, setting K=2, for 50,000 iterations (and 20,000 iterations burn-in). The contribution of λ_1 (measuring the effect of age) to the model fit was assessed by comparing the BIC scores of (i)- the full model, and (ii)- the model in which λ_1 is set to 0.

Results are summarized in Table S4, and show, as expected, that estimates of the effect of age based on unmatched cases and controls are systematically stronger than those based on matched samples. However, in both scenarios, including λ_1 in the model only marginally improved the fit of the model (with differences in BIC scores lower than 10).

This suggests that while estimates of the effect of attained age based on the full population is likely to be underestimated due to age matching, our data do not support an effect of age on the probability to enter lung carcinogenesis, irrespective of age matching.

5 Additional results

Figure S2: Marginal posterior distributions of all parameters (μ^*, $\lambda_1{}^{\ddagger}$, $\lambda_2{}^{\S}$,$\lambda_3{}^{**}$, and $\gamma^{\dagger\dagger}$). Results are presented for K=2.



$^*\mu$: Intercept; $^{\ddagger}\lambda_1$: Effect of age $a^i(t)$; $^{\S}\lambda_2$: Effect of age at starting smoking a_0^i;

$^{**}\lambda_3$: Effect of time since smoking cessation $t_q^i(t)$; $^{\dagger\dagger}\gamma$: Continuous time I_a-I_b transition rate.

Figure S3: Density estimation of the probability to simulate an S to I transition (p_{case}) in actual cases and controls by smoking status. Estimates are based on 10,000 simulated individual trajectories and results are presented for K=2.

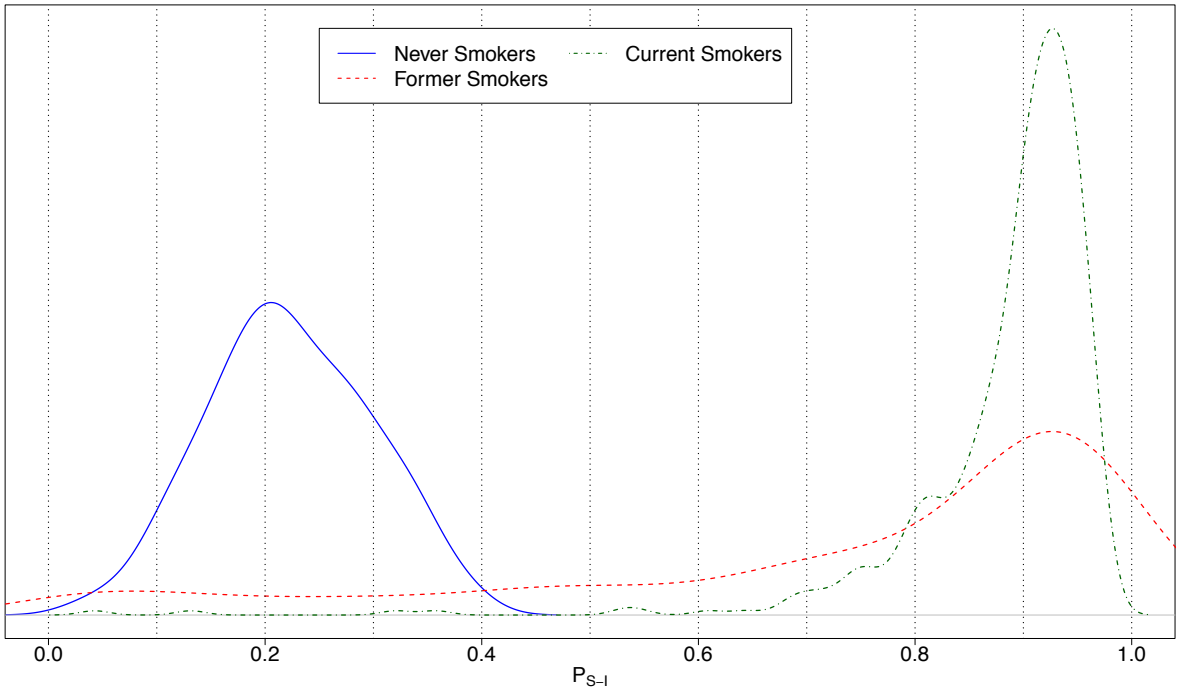


Figure S4: Density estimation of the time to diagnosis. Results for K=2, 5, 10, and 15 are based on 10,000 simulations of individual trajectories derived from the joint posterior distribution of $\mu, \lambda_1, \lambda_2, \lambda_3$, and γ.

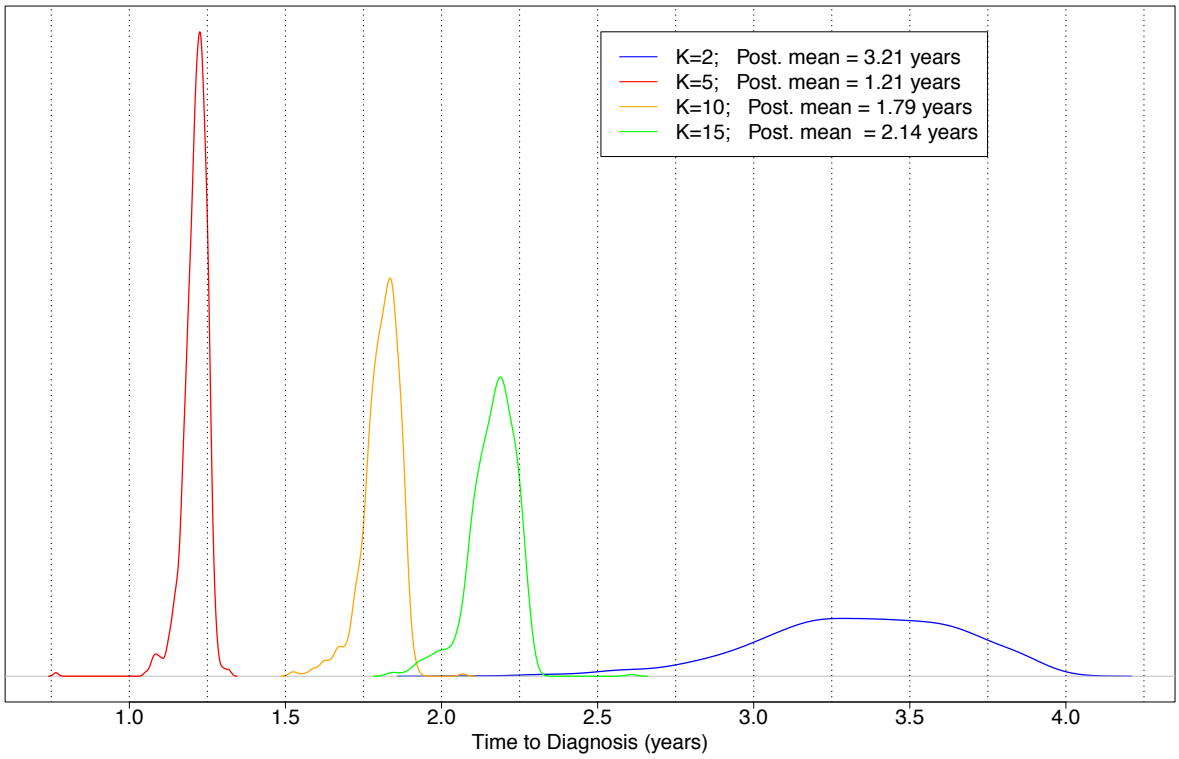


Figure S5: Density estimation of the probability of simulating an S to I transition (p_{case}) in actual cases and controls. Estimates are based on 10,000 simulated individual trajectories and results are presented for K=2, 5, 10 and 15.

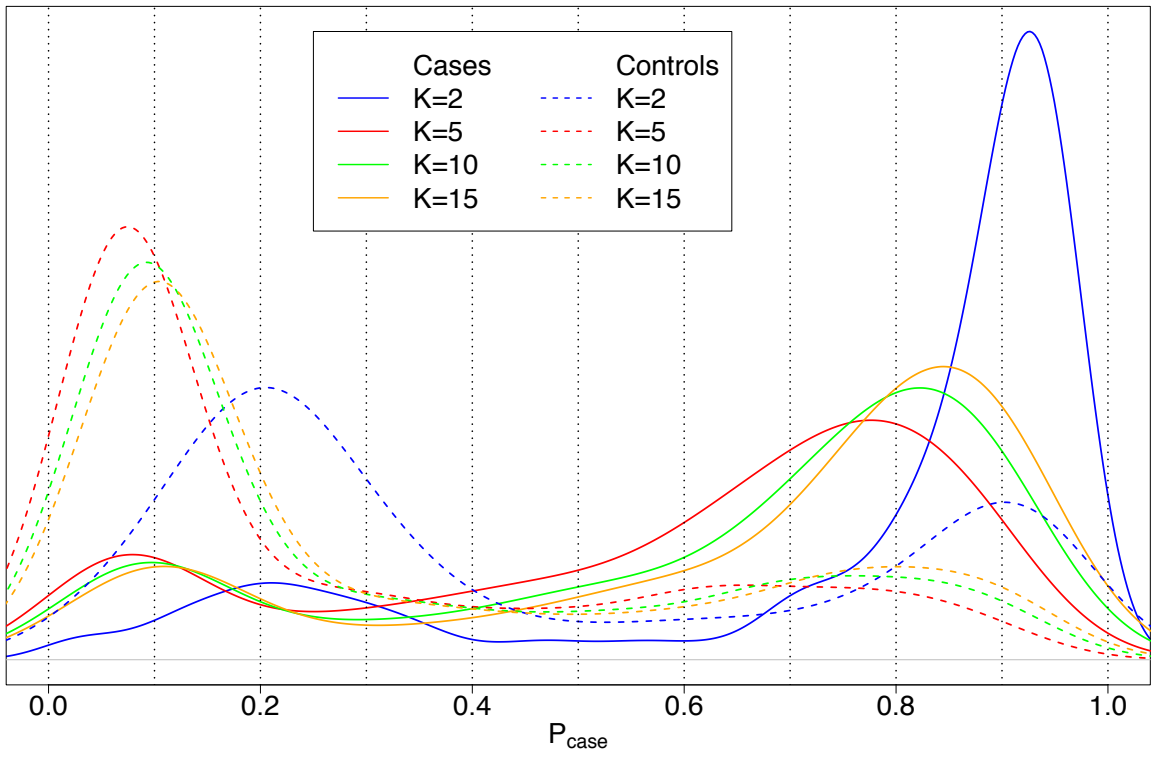
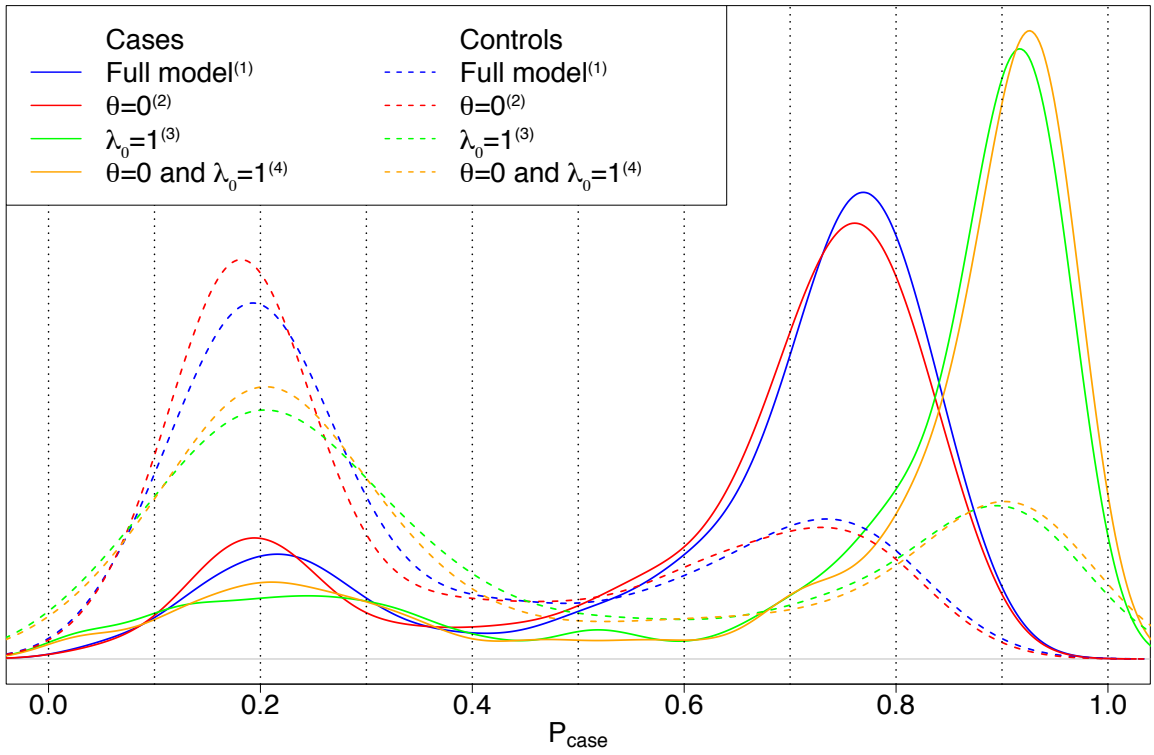


Figure S6: Sensitivity analyses: density estimation of the probability of simulating an S to I transition (p_{case}) in actual cases and controls for the generalised model including an effect of exposure (through λ_0) and an age-dependent sojourn time in I (through θ). Results are based on 50,000 iterations (with 20,000 iterations burn-in), setting K=2.



(1) Full model accounting for age-dependent I_a-I_b transition rates (through θ) and models the effect of exposure (through λ_0)
(2) θ=0 corresponds to a model including the effect of exposure (Eq. (12)) and assuming age-independent I_a-I_b transitions
(3) λ_0=1 corresponds to a model including age-dependent I_a-I_b transitions and assuming a fixed effect of exposure (Eq. (3))
(4) λ_0=1 and θ=0 correspond to the baseline model: age-independent I_a-I_b and fixed effect of exposure (Eq. (3))

Figure S7: Density estimation of the age at inclusion. Results are presented for the cases and controls included in the study and for the full EPIC cohort (N=521,330).

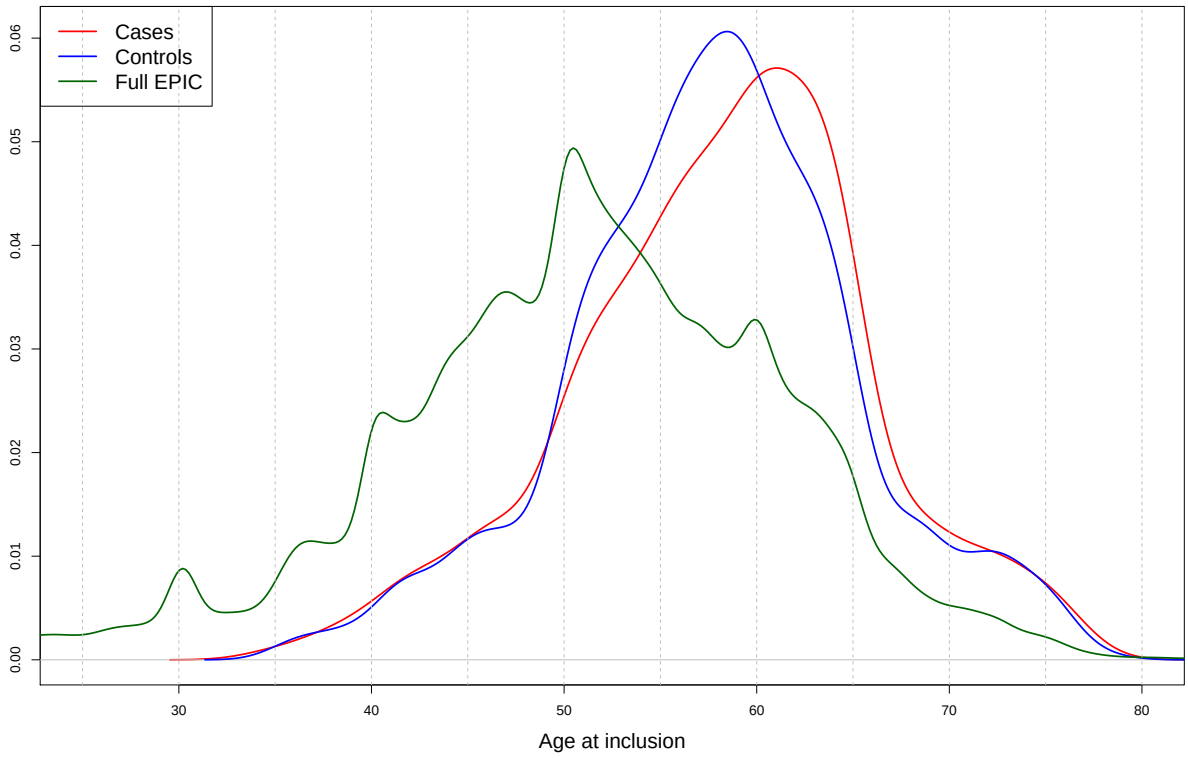


Table S1: Summary features of the study population: case-control data nested in the European Prospective Investigation into Cancer and Nutrition (EPIC) study .

	Smoking Status at enrollment	Sample Size	Blood cotinine level (nmol/L)		
			Mean	Minimum	Maximum
Controls	Never smoker	688	3.7	0.0	73.6
	Former smoker	547	4.2	0.0	85.5
	Current smoker	289	1073.7	5.2	2774.8
	Total	1524	-	-	-
Cases	Never smoker	92	2.2	0.0	17.2
	Former smoker	218	4.4	0.0	69.9
	Current smoker	447	1474.3	212.3	3072.5
	Total	757	-	-	-

Table S2: Sensitivity of parameter estimates to the prior specification. Results are presented for K=2 and based on 50,000 iterations (with 20,000 iterations burn-in).

Prior Distribution	Parameter Estimates: posterior mean (95% Credible Intervals)				
	$\mu^{(1)}$	${\lambda_1}^{(2)}$	${\lambda_2}^{(3)}$	${\lambda_3}^{(4)}$	$\gamma^{(5)}$
${\mathcal{U}_{[-100;100]}}^{(6)}$	1.52 (0.44;2.86)	0.03 (0.01;0.05)	-0.08(-0.10;-0.06)	-0.06 (-0.08;-0.05)	2.45 (2.13;2.81)
$\mathcal{N}(0,1000)^{(7)}$	1.51 (0.33; 2.73)	0.03 (0.01; 0.05)	-0.08 (-0.10; -0.06)	-0.06 (-0.08; -0.05)	2.45 (2.12; 2.81)
$\mathcal{N}(0,100)^{(7)}$	1.60 (0.45; 2.74)	0.03 (0.01; 0.05)	-0.08 (-0.10; -0.06)	-0.06 (-0.08; -0.05)	2.46 (2.11; 2.82)
$\mathcal{N}(0,10)^{(7)}$	1.51 (0.33; 2.73)	0.03 (0.01; 0.05)	-0.08 (-0.10; -0.06)	-0.06 (-0.08; -0.05)	2.45 (2.12; 2.81)

$^{(1)}\mu$: Intercept
$^{(2)}\lambda_1$: Effect of age $a^i(t)$
$^{(3)}\lambda_2$: Effect of age at starting smoking a_0^i
$^{(4)}\lambda_3$: Effect of time since smoking cessation $t_q^i(t)$
$^{(5)}\gamma$: continuous time I_i-I_j transition rate
$^{(6)}\mathcal{U}_{[-100;100]}$: uniform distribution with support [-100;100]
$^{(7)}\mathcal{N}(0,\sigma^2)$: zero-centered Gaussian distribution with variance σ^2

Table S3: Sensitivity analyses: parameter estimates for the generalised model including the effect of exposure (through λ_0) and an age-dependent sojourn time in I (through θ). Results are based on 50,000 iterations (with 20,000 iterations burn-in), setting K=2.

	Parameters Estimates: posterior mean (95% Credible Intervals)							BIC
	$\mu^{(1)}$	${\lambda_0}^{(2)}$	${\lambda_1}^{(3)}$	${\lambda_2}^{(4)}$	${\lambda_3}^{(5)}$	$\theta^{(6)}$	${\gamma_0}^{(7)}$	
Full model[8]	-2.59 (-3.49;-1.55)	0.51 (0.44;0.59)	0.04 (0.02;0.05)	-0.01 (-0.03;0.01)	-0.05 (-0.06;-0.04)	0.02 (0.01;0.03)	0.52 (0.25;0.86)	8166.6
θ=0[9]	-2.01 (-2.94;-1.09)	0.5 (0.42;0.58)	0.03 (0.02;0.04)	0 (-0.02;0.02)	-0.05 (-0.06;-0.04)	-	1.29 (1.01;1.57)	8174.0
λ_0=1[10]	0.57 (-0.7;1.71)	-	0.04 (0.03;0.06)	-0.09 (-0.11;-0.07)	-0.06 (-0.08;-0.04)	0.01 (0.01;0.02)	1.21 (0.69;1.78)	8247.5
θ=0; λ_0=1[11]	1.52 (0.44;2.86)	-	0.03 (0.01;0.05)	-0.08 (-0.1;-0.06)	-0.06 (-0.08;-0.05)	-	2.45 (2.13;2.81)	8255.6

[1] μ : Intercept
[2] λ_0 : Main effect of exposure
[3] λ_1 : Effect of age $a^i(t)$
[4] λ_2 : Effect of age at starting smoking a_0^i
[5] λ_3 : Effect of time since smoking cessation $t_q^i(t)$
[6] θ : Linear effect (on log scale) of age on the I_i-I_j transition rate
[7] γ_0 : background continuous time I_i-I_j transition rate
[8] Full model accounting for age-dependent I_a-I_b transition rates (through θ and modelling the effect of exposure (through λ_0)
[9] Setting θ=0 corresponds to a model including the effect of exposure (Eq. (12)) and assuming age-independent I_a-I_b transitions
[10] Setting λ_0=1 corresponds to a model including age-dependent I_a-I_b transitions and assuming a fixed effect of exposure (Eq. (3))
[11] Setting λ_0=1 and θ=0 corresponds to a model assuming age-independent I_a-I_b transitions and a fixed effect of exposure (Eq. (3)), as presented in the main text

Table S4: Posterior mean (05% credible interval) of λ_1, measuring the effect of age for 2 sub-samples each containing all cases diagnosed more than one year after enrolement (N=738) and (N=738) controls. Controls were resampled with and without age matching. Results are presented for K=2 and based on 50,000 iterations (with 20,000 iterations burn-in).

	With age matching		Without age matching	
Sample	$\lambda_1^{(1)}$	$\Delta(BIC)^{(2)}$	$\lambda_1^{(1)}$	$\Delta(BIC)^{(2)}$
1	0.009 (-0.03;0.04)	7.54	0.030 (-0.01;0.07)	5.26
2	0.009 (-0.03;0.05)	7.04	0.027 (-0.01;0.06)	5.68
3	0.003 (-0.04;0.04)	7.72	0.024 (-0.01;0.05)	4.50
4	0.004 (-0.03;0.04)	7.72	0.010 (-0.02;0.05)	7.46
5	0.016 (-0.02;0.05)	7.34	0.027 (0.00;0.06)	5.60
6	0.003 (-0.03;0.04)	7.74	0.033 (0.00;0.07)	3.26
7	0.002 (-0.03;0.04)	7.62	0.035 (0.01;0.06)	3.54
8	0.003 (-0.03;0.03)	7.66	0.024 (-0.01;0.06)	5.36
9	-0.001 (-0.04;0.04)	7.74	0.032 (0.00;0.06)	3.34
10	0.004 (-0.03;0.04)	7.58	0.030 (-0.01;0.06)	4.70
11	0.005 (-0.03;0.04)	7.50	0.035 (0.01;0.07)	3.14
12	0.007 (-0.03;0.04)	7.40	0.024 (-0.01;0.06)	5.38
13	0.008 (-0.03;0.04)	6.72	0.048 (0.02;0.08)	0.24
14	-0.001 (-0.04;0.04)	7.76	0.034 (0.00;0.06)	3.28
15	0.002 (-0.04;0.04)	7.70	0.033 (0.00;0.07)	4.42
16	0.008 (-0.03;0.04)	7.58	0.040 (0.01;0.07)	1.56
17	0.003 (-0.03;0.04)	7.68	0.031 (0.00;0.06)	5.00
18	0.014 (-0.02;0.04)	7.12	0.043 (0.01;0.08)	0.62
19	-0.002 (-0.03;0.03)	7.76	0.026 (-0.01;0.06)	5.32
20	0.004 (-0.03;0.04)	7.76	0.021 (-0.02;0.06)	6.82

$^{(1)}\lambda_1$: Effect of age $a^i(t)$

$^{(2)}\Delta(BIC)$: Difference in the Bayesian Information Criterion (BIC) scores for the model in which λ_1 is estimated and the one where λ_1=0