Supplementary material

Model derivation

This derivation follows that of Brouwer [S1]. Let X, $Y(t)$, $Z(t)$, $P(t)$, ν_0, α, β, μ_1, and σ be as described in Table 1. Let the per-cell initiation rate be written as $\nu(t) = \nu_0 + \sigma P(t)$.

Write the following generating equations for age $t \geq \tau$.

$$\Psi(y, z, \tau, t) = E[y^{Y(t)} z^{Z(t)} | Y(\tau) = 0, Z(\tau) = 0)], \tag{S1}$$

$$\Phi(y, z, \tau, t) = E[y^{Y(t)} z^{Z(t)} | Y(\tau) = 1, Z(\tau) = 0], \tag{S2}$$

$$\Theta(y, z, \tau, t) = E[y^{Y(t)} z^{Z(t)} | Y(\tau) = 0, Z(\tau) = 1]. \tag{S3}$$

In terms of the probability generating function, the survival and hazard functions are as follows.

$$S(t) = \sum_j P_{(0,0),(j,0)}(0, t) = \sum_{j,k} P_{(0,0),(j,k)}(0, t) 1^j 0^k = \Psi(1, 0, t), \tag{S4}$$

$$h(t) = -\frac{\Psi'(1, 0, t)}{\Psi(1, 0, t)}. \tag{S5}$$

We may write the Kolmogorov forward equations for the probability generating functions as follows, suppressing dependence on τ and t

$$\begin{aligned} \frac{\partial \Psi}{\partial \tau} &= \nu(\tau) X (1 - \Phi) \Psi, \\ \frac{\partial \Phi}{\partial \tau} &= [\alpha + \beta + \mu_1] \Phi - \beta - \alpha \Phi^2 - \mu_1 \Phi \Theta, \\ \frac{\partial \Theta}{\partial \tau} &= 0, \end{aligned} \tag{S6}$$

with initial conditions $\Psi(y, z, t, t) = 1$, $\Phi(y, z, t, t) = y$, and $\Theta(y, z, t, t) = z$. Denote derivative with respect to t as $'$. Then

$$\begin{aligned} \frac{\partial \Psi'}{\partial \tau} &= -\nu(\tau) X (\Phi' \Psi + (\Phi - 1) \Psi'), \\ \frac{\partial \Phi'}{\partial \tau} &= [\alpha + \beta + \mu_1] \Phi' - 2\alpha \Phi \Phi' - \mu_1 (\Phi' \Theta + \Phi \Theta'), \\ \frac{\partial \Theta'}{\partial \tau} &= 0, \end{aligned} \tag{S7}$$

with initial conditions $\Psi'(y, z, t{-}0, t) = -\mu_0(t) X(t) (1{-}y)$, $\Phi'(y, z, t{-}0, t) = -[\alpha(t) + \beta(t) + \mu_1(t)]\, y + \beta(t) + \alpha(t) y^2 + \mu_1(t) yz$, and $\Theta'(y, z, t{-}0, t) = 0$. From these equations, it is clear that $\Theta'(y, z, \tau, t) \equiv 0$ and $\Theta(y, z, \tau, t) = z$.

Let $\Gamma(y,z,\tau,t) = -\ln\Psi(y,z,\tau,t)$, so that

$$\begin{aligned}\frac{\partial\Gamma}{\partial\tau} &= -\nu(\tau)X(\tau)\left(1-\Phi\right),\\ \frac{\partial\Gamma'}{\partial\tau} &= \nu(\tau)X(\tau)\Phi'.\end{aligned} \tag{S8}$$

Let $x_1(s) = \Psi(1,0,t-s,t)$, and $x_2(s) = \Gamma'(1,0,t-s,t)$, $x_3(s) = \Phi(1,0,t-s,t)$, and $x_4(s) = \Phi'(1,0,t-s,t)$, and write the following system of equations.

$$\begin{aligned}\frac{\partial x_1}{\partial s}(s) &= -\nu(t-s)Xx_1(1-x_3),\\ \frac{\partial x_2}{\partial s}(s) &= -\nu(t-s)Xx_4,\\ \frac{\partial x_3}{\partial s}(s) &= -\left[\alpha+\beta+\mu_1\right]x_3+\beta+\alpha x_3^2,\\ \frac{\partial x_4}{\partial s}(s) &= -\left[\alpha+\beta+\mu_1\right]x_4+2\alpha x_3x_4,\end{aligned} \tag{S9}$$

with initial conditions $x_1(0) = 1$, $x_2(0) = 0$, $x_3(0) = 1$, $x_4(0) = -\mu_1$. Solving this set of equations for each value of t, we recover the survival $S(t) = x_1(t)$ and hazard $h(t) = x_2(t)$ functions. The hazard function $h(t)$ corresponds to the age-specific cancer incidence data. Although there is a closed-form solution to the hazard for constant parameters, when $\nu(t)$ is not constant (i.e., when $P(t)$ is not constant), the hazards must be solved for numerically.

The three-stage model derivation follows that of the the two-stage model and gives the following system of equations. Again, we may solve for the survival $S(t) = x_1(t)$ and hazard $h(t) = x_2(t)$ functions.

$$\begin{aligned}\frac{\partial x_1}{\partial s}(s) &= -\nu(t-s)Xx_1(1-x_3),\\ \frac{\partial x_2}{\partial s}(s) &= -\nu(t-s)Xx_4,\\ \frac{\partial x_3}{\partial s}(s) &= -\mu_1x_3(1-x_5),\\ \frac{\partial x_4}{\partial s}(s) &= -\mu_1x_4(1-x_5)+\mu_1x_3x_6,\\ \frac{\partial x_5}{\partial s}(s) &= -\left[\alpha+\beta+\mu_2\right]x_3+\beta+\alpha x_5^2,\\ \frac{\partial x_6}{\partial s}(s) &= -\left[\alpha+\beta+\mu_2\right]x_6+2\alpha x_5x_6,\end{aligned} \tag{S10}$$

with initial conditions $x_1(0) = 1$, $x_2(0) = 0$, $x_3(0) = 1$, $x_4(0) = 0$, $x_5(0) = 1$, $x_6(0) = -\mu_2$.

Additional figures

Here, we provide figures of the prevalence $P(t)$ used for *H. pylori*, smoking, and cervicogenital HPV (Figure S1), the cancer incidence model predictions for each cancer (Figure S2), and the model fits to gastric cancer incidence when *H. pylori* prevalence is estimated rather than assumed (Figure S3), and the birth cohort effects for HPV prevalence estimated from HPV-related cancer incidence (Figure S4).

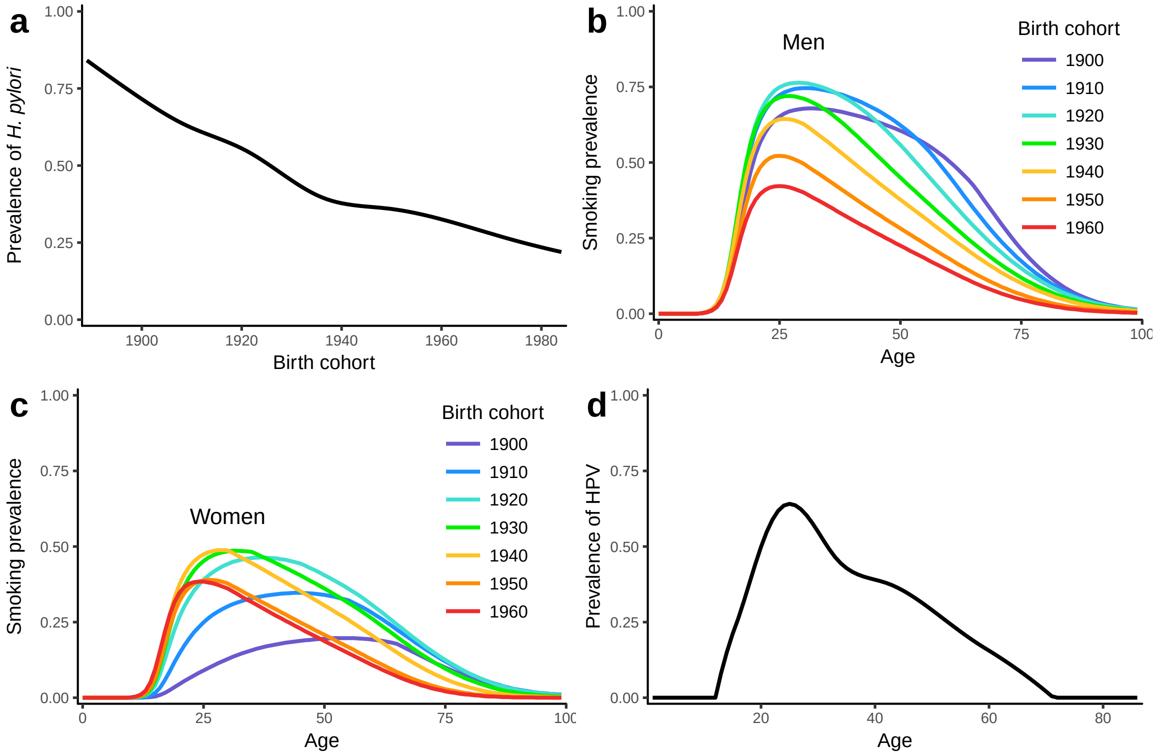

Figure S1: Prevalence $P(t)$ of a) *H. pylori*, smoking for b) men and c) women, and d) cervicogenital human papillomavirus.

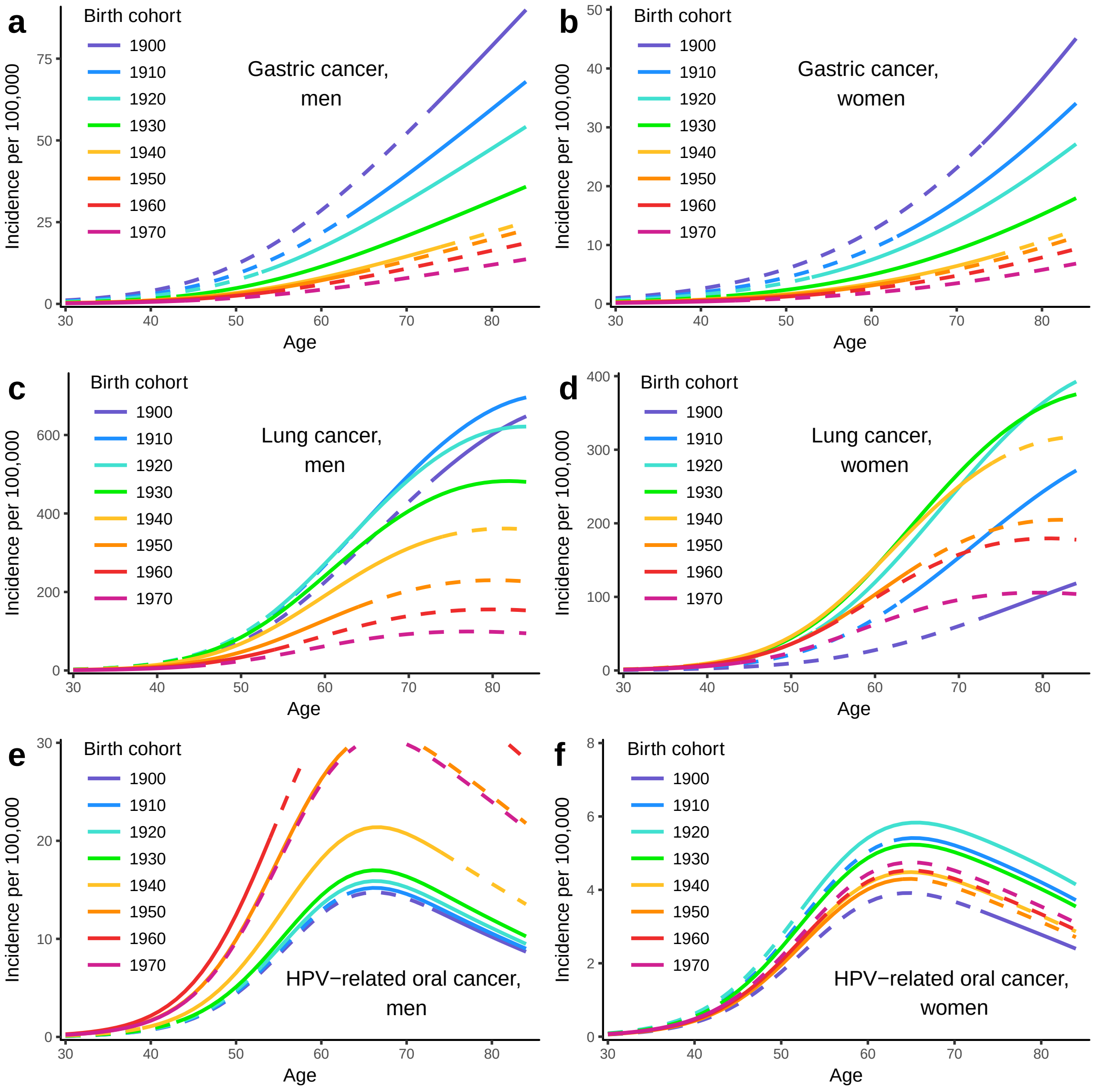

Figure S2: Modeled and predicted incidence per 100,000 for men and women by birth cohort of (a and b) intestinal-type noncardia gastric adenocarcinoma (GC), (c and d) malignant neoplasms of the bronchus and lung (LC), and (e anf f) HPV-related oral (oropharyngeal and oral cavity) squamous cell carcinoma (OSCC). Solid lines are the model hazards corresponding to years where there is data, and the dotted lines are the model hazards for years without corresponding data.

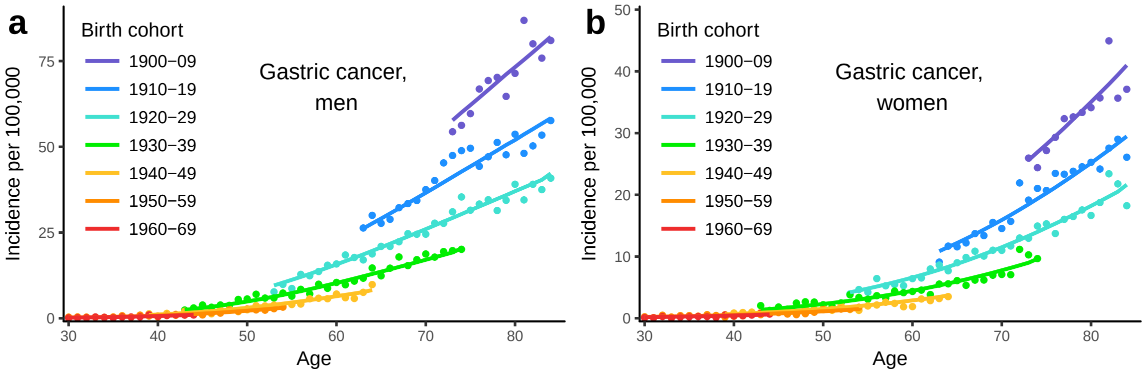

Figure S3: Incidence and modeled incidence per 100,000 of intestinal-type noncardia gastric adenocarcinoma by birth cohort for men and women when also fitting relative *H. pylori* prevalence. Dots are SEER 9 data, and the lines are the model hazards.

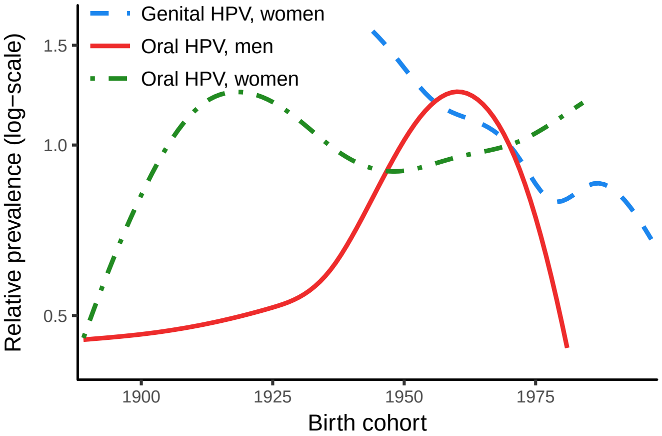

Figure S4: Modeled relative prevalence of oral and genital HPV by birth cohort (relative to the 1970 birth cohort). Oral HPV cohort effects are estimated from the HPV-related oral cancer incidence, and genital HPV cohort effects are derived from Brouwer et al. [S2]

Etiological agent driving promotion

We considered a model where the prevalence of the etiological agent affected promotion rather than initiation rate. We parameterized this model by adjusting the net cell proliferation $\alpha - \beta - \mu_1$ by a factor $1 + \varphi P(t)$. This parameterization can be written in the following system of differential equations (as seen in the computation of the identifiable parameter combinations), with the variables defined as before.

$$\begin{aligned}
&\frac{\partial x_1}{\partial s}(s) = -\nu_0(t-s)Xx_1(1-x_3)\\
&\frac{\partial x_2}{\partial s}(s) = -\nu_0(t-s)Xx_4\\
&\frac{\partial x_3}{\partial s}(s) = -\left[\alpha+\beta+\mu_1+[-\alpha+\beta+\mu_1][\varphi P(t-s)]\right]x_3 + [\beta+[-\alpha+\beta+\mu_1]\varphi P(t-s)] + \alpha x_3^2\\
&\frac{\partial x_4}{\partial s}(s) = -\left[\alpha+\beta+\mu_1\right]x_4 + 2\alpha x_3 x_4.
\end{aligned} \tag{S11}$$

The fits of this model to gastric and lung cancer are given in Figure S5. In all but one case, the model with effects on initiation fit better than with effects on promotion (difference in log-likelihood: gastric cancer for men (-18.5), gastric cancer for women (9.8), lung cancer for men (689.7), lung cancer for women (4457.2)). Parameter estimates are given in Table S1.

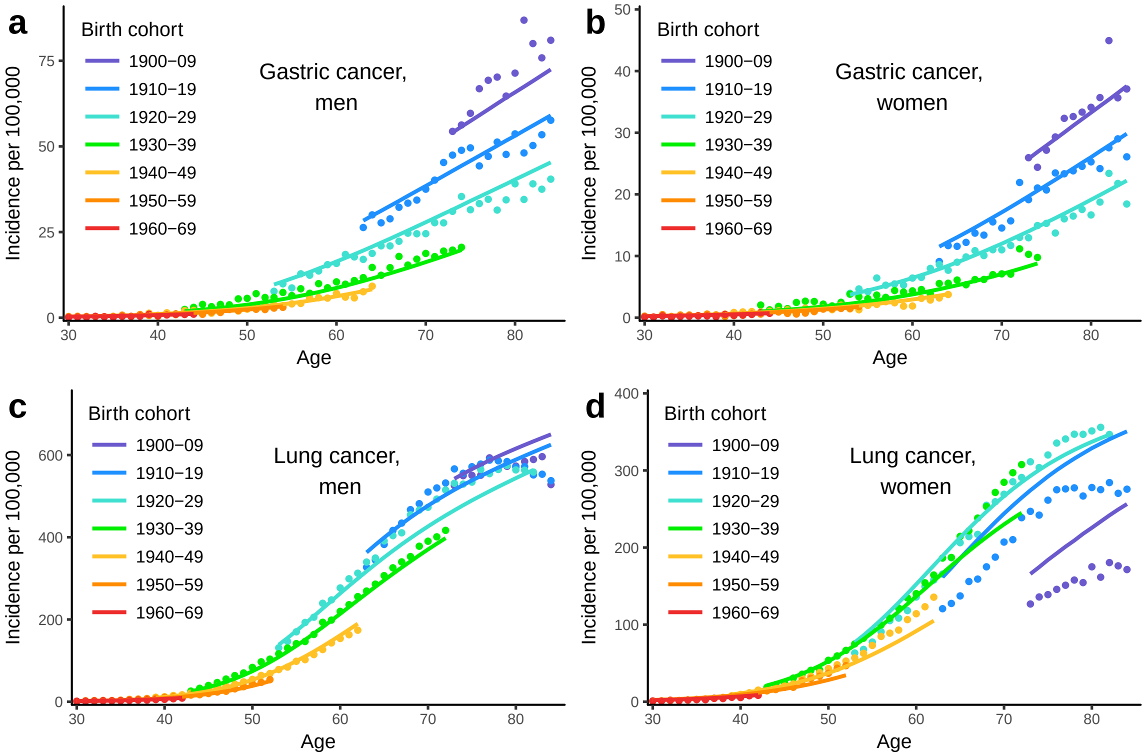

Figure S5: Modeled and predicted incidence per 100,000 for men and women by birth cohort of (a and b) intestinal-type noncardia gastric adenocarcinoma (GC), (c and d) malignant neoplasms of the bronchus and lung (LC) using the model with etiological agent effects on promotion.

Table S1: Estimated carcinogenesis parameters for gastric, lung, and HPV-related oral cancer. The number of stages used for the multistage clonal expansion model used and whether the etiological agent prevalence is assumed (forward framework) or fit (inverse framework) are indicated.

Cancer	Model	Prevalence	Sex	p (95% CI)	q (95% CI)	r (95% CI)	φ (95% CI)
Gastric	3-stage	Assumed	Male	-0.009 (-0.012, -0.006)	$2.67\cdot10^{-3}$ ($1.84\cdot10^{-3}$, $3.84\cdot10^{-3}$)	$9.83\cdot10^{-3}$ ($9.41\cdot10^{-3}$, $1.03\cdot10^{-2}$)	41.20 (19.44, 67.04)
			Female	-0.011 (-0.018, -0.010)	$2.34\cdot10^{-3}$ ($1.35\cdot10^{-3}$, $4.05\cdot10^{-3}$)	$9.31\cdot10^{-3}$ ($8.71\cdot10^{-3}$, $9.96\cdot10^{-3}$)	19.28 (8.72, 37.38)
Lung	3-stage	Assumed	Male	-0.122 (-0.126, -0.122)	$1.86\cdot10^{-5}$ ($1.83\cdot10^{-5}$, $1.90\cdot10^{-5}$)	$3.86\cdot10^{-2}$ ($3.84\cdot10^{-3}$, $3.87\cdot10^{-2}$)	0.960 (0.954, 0.965)
			Female	-0.071 (-0.071, -0.071)	$8.19\cdot10^{-5}$ ($8.10\cdot10^{-5}$, $8.28\cdot10^{-5}$)	$3.95\cdot10^{-2}$ ($3.93\cdot10^{-2}$, $3.98\cdot10^{-2}$)	2.79 (2.77, 2.82)

Identifiability

The proofs of the theoretical results in this section follow the method of proof in [S3].

Proposition 1. *If cancer survival (or, equivalently, age-specific incidence) and etiological agent prevalence $P(t)$ are perfectly measured, the two-stage clonal expansion model with background and agent initiation and constant parameters (ν_0, σ, X, α, β, μ_1) is unidentifiable but has four identifiable parameter combinations, which may be represented as $\nu_0 X \mu_1$, $\sigma X \mu_1$, $\alpha\mu_1$, and $\alpha - \beta - \mu_1$.*

Proof. Let $u(s) = P(t-s)$. Assume that the cancer survival, x_1, is perfectly measured. The aim of this proof is to determine the structurally identifiable parameter combinations of the two-stage model model with background and agent initiation by observing the coefficients of the input–output equation for the system, which is a monic polynomial of the observed input u, output x_1 and their derivatives.

Then the following equations contain all information of the two-stage clonal expansion model (because we hazard and survival contain equivalent information, it is sufficient to use only the equations related to the survival).

$$\dot{x}_1 = -(\nu_0 + \sigma u) X x_1 (1 - x_3), \tag{S12}$$

$$\dot{x}_3 = -\left[\alpha + \beta + \mu_1\right] x_3 + \beta + \alpha x_3^2. \tag{S13}$$

We solve for x_3 in terms of x_1 and its derivatives,

$$x_3 = \frac{x_1 + \frac{\dot{x}_1}{\nu_0 + \sigma u}}{x_1}. \tag{S14}$$

Plug this in to the $\dot{x}_1$ equation (Eq. (S24)),

$$\left(\frac{x_1 + \frac{\dot{x}_1}{\nu_0 X + \sigma X u}}{x_1}\right)' = -(\alpha + \beta + \mu_1)\left(\frac{x_1 + \frac{\dot{x}_1}{\nu_0 X + \sigma X u}}{x_1}\right) + \beta + \alpha\left(\frac{x_1 + \frac{\dot{x}_1}{\nu_0 X + \sigma X u}}{x_1}\right)^2, \tag{S15}$$

simplifying to

$$\begin{aligned} 0 = {} & u x_1 \ddot{x}_1 - \dot{u} x_1 \dot{x}_1 - u \dot{x}_1^2 - (\alpha - \beta - \mu_1) u x_1 \dot{x}_1 + 2(\mu_1 \nu_0 X) u x_1^2 + (\mu_1 \sigma X) u^2 x_1^2 \\ & + \frac{\mu_1 \nu_0 X}{\mu_1 \sigma X} x_1 \ddot{x}_1 - (\alpha - \beta - \mu_1)\frac{\mu_1 \nu_0 X}{\mu_1 \sigma X} x_1 \dot{x}_1 + \frac{(\mu_1 \nu_0 X)^2}{\mu_1 \sigma X} x_1^2 - \frac{(\alpha\mu_1 + \mu_1 \nu_0 X)}{\mu_1 \sigma X} \dot{x}_1^2, \end{aligned} \tag{S16}$$

This last equation is a monic polynomial of x_1, u, and their derivatives, is equivalent to the original differential equations, and is thus an input–output equation. We can read a set of identifiable parameter combinations from the equation coefficients: $\mu_1\nu_0X$, $\mu_1\sigma X$, $\alpha-\beta-\mu_1$, and $\alpha\mu_1$. An equivalent set of identifiable combinations is

$$p,q = \frac{1}{2}\left(-(\alpha-\beta-\mu_1) \mp \sqrt{(\alpha-\beta-\mu_1)^2+4\alpha\mu_1}\right) \tag{S17}$$

$$r = \frac{\nu_0 X}{\alpha}, \tag{S18}$$

$$s = \frac{\sigma X}{\alpha}. \tag{S19}$$

□

Corollary 1. *If cancer survival (or, equivalently, age-specific incidence) is perfectly measured and agent prevalence is to be estimated, the two-stage clonal expansion model with background and agent initiation and constant parameters (ν_0, σ, X, α, β, μ_1) is unidentifiable but has four identifiable parameter combinations, which may be represented as $\nu_0X\mu_1$, $\sigma X\mu_1P(t)$, $\alpha\mu_1$, and $\alpha-\beta-\mu_1$.*

Proof. Here, we assume that u is rational (or approximable as rational). Writing the input–output equation in the following form, we see that σX is not identifiable separately from u:

$$\begin{aligned} 0 = x_1\ddot{x}_1 &- \left[\frac{\mu_1\sigma X\dot{u}}{\mu_1\sigma Xu} + (\alpha-\beta-\mu_1)\left(\frac{\mu_1\nu_0X}{\mu_1\sigma Xu}+1\right)\right]x_1\dot{x}_1 + \frac{\mu_1\nu_0X}{\mu_1\sigma Xu}x_1\ddot{x}_1 \\ &+ \frac{(\mu_1\nu_0X+\mu_1\sigma Xu)^2}{\mu_1\sigma Xu}x_1^2 - \left(\frac{(\alpha\mu_1+\mu_1\nu_0X)}{\mu_1\sigma Xu}+1\right)\dot{x}_1^2. \end{aligned} \tag{S20}$$

Since $u(s) = P(t-s)$, $\sigma X\mu_1P(t)$ is identifiable for each t. □

Proposition 2. *If cancer survival (or, equivalently, age-specific incidence) and etiological agent prevalence $P(t)$ are perfectly measured, the three-stage clonal expansion model with background and agent initiation and constant parameters (ν_0, σ, X, μ_1, μ_2, α, β) is unidentifiable but has five identifiable parameter combinations, which may be represented as ν_0X, σX, $\mu_1\mu_2$, $\alpha_1\mu_2$, and $\alpha-\beta-\mu_2$.*

The proof is analogous to that of the two-stage model. Although the five combinations above are structurally identifiable, we previously demonstrated that the combinations involving the intermediate mutation rates were not practically identifiable in the three-stage model [S4]. Hence, we use the following set of practically identifiable combinations

$$p,q = \frac{1}{2}\left(-(\alpha-\beta-\mu_2) \mp \sqrt{(\alpha-\beta-\mu_2)^2+4\alpha\mu_2}\right), \tag{S21}$$

$$r = \sqrt{\nu_0\mu_1X/\alpha}, \tag{S22}$$

$$s = \sigma\sqrt{X/\alpha}. \tag{S23}$$

Proposition 3. *If cancer survival (or, equivalently, age-specific incidence) and etiological agent prevalence $P(t)$ are perfectly measured, the two-stage clonal expansion model with background and agent promotion (Eqs.* (S11)*) and constant parameters (ν_0, σ, X, α, β, μ_1) is unidentifiable but has four identifiable parameter combinations, which may be represented as $\nu_0 X\mu_1$, φ, $\alpha\mu_1$, and $\alpha - \beta - \mu_1$.*

Proof. Let $u(s) = P(t - s)$. Assume that the cancer survival, x_1, is perfectly measured. The aim of this proof is to determine the structurally identifiable parameter combinations of the two-stage model model with background and agent initiation by observing the coefficients of the input–output equation for the system, which is a monic polynomial of the observed input u, output x_1 and their derivatives.

Then the following equations contain all information of the two-stage clonal expansion model (because we hazard and survival contain equivalent information, it is sufficient to use only the equations related to the survival).

$$\dot{x}_1 = -(\nu_0)Xx_1(1 - x_3), \tag{S24}$$

$$\dot{x}_3 = -\left[\alpha + \beta + \mu_1 + [-\alpha + \beta + \mu_1][\varphi u]\right]x_3 + [\beta + [-\alpha + \beta + \mu_1]\varphi u] + \alpha x_3^2. \tag{S25}$$

We solve for x_3 in terms of x_1 and its derivatives,

$$x_3 = \frac{x_1 + \frac{\dot{x}_1}{\nu_0}}{x_1}. \tag{S26}$$

Plug this in to the $\dot{x}_1$ equation (Eq. (S24)),

$$\left(\frac{x_1 + \frac{\dot{x}_1}{\nu_0 X}}{x_1}\right)' = -\left[\alpha + \beta + \mu_1 + [-\alpha + \beta + \mu_1][\varphi u]\right]\left(\frac{x_1 + \frac{\dot{x}_1}{\nu_0 X}}{x_1}\right) \tag{S27}$$

$$+ [\beta + [-\alpha + \beta + \mu_1]\varphi u] + \alpha\left(\frac{x_1 + \frac{\dot{x}_1}{\nu_0 X}}{x_1}\right)^2, \tag{S28}$$

simplifying to

$$\begin{aligned} 0 = & x_1\ddot{x}_1 + 2\nu_0 X x_1\dot{x}_1 + \mu_1\nu_0 X x_1^2 - \left(\frac{\alpha\mu_1}{\mu_1\nu_0 X} - 1\right)(\dot{x}_1)^2 \\ & - (\alpha - \beta - \mu_1)x_1\dot{x}_1 - (\alpha - \beta - \mu_1)\varphi u x_1\dot{x}_1 \end{aligned} \tag{S29}$$

This last equation is a monic polynomial of x_1, u, and their derivatives, is equivalent to the original differential equations, and is thus an input–output equation. We can read a set of identifiable parameter combinations from the equation coefficients: $\mu_1\nu_0 X$, φ, $\alpha{-}\beta{-}\mu_1$, and $\alpha\mu_1$. An equivalent set of identifiable combinations is

$$p, q = \frac{1}{2}\left(-(\alpha-\beta-\mu_1) \mp \sqrt{(\alpha-\beta-\mu_1)^2 + 4\alpha\mu_1}\right) \tag{S30}$$

$$r = \frac{\nu_0 X}{\alpha}, \tag{S31}$$

$$\varphi. \tag{S32}$$

□

References

[S1] Brouwer AF. Models of HPV as an Infectious Disease and as an Etiological Agent of Cancer. University of Michigan; 2015.

[S2] Brouwer AF, Eisenberg MC, Carey TE, Meza R. Trends in HPV cervical and seroprevalence and associations between oral and genital infection and serum antibodies in NHANES 2003–2012. BMC Infectious Diseases. 2015;15(1):575.

[S3] Brouwer AF, Meza R, Eisenberg MC. A systematic approach to determining the identifiability of multistage carcinogenesis models. Risk Analysis. 2016;.

[S4] Brouwer AF, Meza R, Eisenberg MC. Parameter estimation for multistage clonal expansion models from cancer incidence data: a practical identifiability analysis. PLOS Computational Biology. 2017;.