PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis

**Andrew F. Brouwer*, Rafael Meza, Marisa C. Eisenberg**

Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, United States of America

* brouweaf@umich.edu





**Citation:** Brouwer AF, Meza R, Eisenberg MC (2017) Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. PLoS Comput Biol 13(3): e1005431. https://doi.org/10.1371/journal.pcbi.1005431

**Editor:** Rachel Karchin, Johns Hopkins University, UNITED STATES

**Received:** September 27, 2016

**Accepted:** February 25, 2017

**Published:** March 13, 2017



**Data Availability Statement:** All relevant data are within the paper and its Supporting Information files.

**Funding:** This work was supported by NIH (https://www.nih.gov/) grant U01CA182915. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

Many cancers are understood to be the product of multiple somatic mutations or other rate-limiting events. Multistage clonal expansion (MSCE) models are a class of continuous-time Markov chain models that capture the multi-hit initiation–promotion–malignant-conversion hypothesis of carcinogenesis. These models have been used broadly to investigate the epidemiology of many cancers, assess the impact of carcinogen exposures on cancer risk, and evaluate the potential impact of cancer prevention and control strategies on cancer rates. Structural identifiability (the analysis of the maximum parametric information available for a model given perfectly measured data) of certain MSCE models has been previously investigated. However, structural identifiability is a theoretical property and does not address the limitations of real data. In this study, we use pancreatic cancer as a case study to examine the practical identifiability of the two-, three-, and four-stage clonal expansion models given age-specific cancer incidence data using a numerical profile-likelihood approach. We demonstrate that, in the case of the three- and four-stage models, several parameters that are theoretically structurally identifiable, are, in practice, unidentifiable. This result means that key parameters such as the intermediate cell mutation rates are not individually identifiable from the data and that estimation of those parameters, even if structurally identifiable, will not be stable. We also show that products of these practically unidentifiable parameters are practically identifiable, and, based on this, we propose new reparameterizations of the model hazards that resolve the parameter estimation problems. Our results highlight the importance of identifiability to the interpretation of model parameter estimates.

## Author summary

Parameter estimation from data is an important part of mathematical modeling, and structural identifiability is the study of what parametric information exists, for a given model, in ideal data. Unfortunately, for a variety of reasons, there is often less information available in our real data sets. The study of these problems is called practical identifiability. In this study, we consider a family of models of cancer biology that are commonly used to explain cancer incidence in terms of underlying biological parameters. Using profile

likelihoods, a widely applicable numerical tool, we demonstrate that even though the more complex models we consider have theoretically more identifiable parameters, the data contains only three pieces of practically identifiable information for each model: the product of the initiating mutation rates, the net cell proliferation rate, and the scaled malignant conversion rate. This result can be interpreted biologically: we can determine only the product of cell mutation rates not the intermediate rates themselves. Our result limits the interpretability of previous work, but we propose a novel parameterization to resolve the identifiability issue. Ultimately, our analysis demonstrates the importance of verifying the practical identifiability of parameters before assigning too much weight to the interpretation of their estimated values.

## Introduction

Parameter estimation is an important aspect of computational modeling in the life sciences because parameter estimates can shed light on underlying biological mechanisms and processes and provide a way to link dynamic models to real-world data. However, the dynamics of many living systems have evolved to be robust to changes in underlying parameters, which necessitates an understanding of which parameters or combinations of parameters can even be estimated from data, known as identifiability. Here, we leverage computational identifiability tools to determine what cancer incidence data can tell us about the biology of carcinogenesis.

Cancers arise from the accumulation of genetic (and epigenetic) abnormalities and mutations. Although a single change is thought to be sufficient for certain cancers (certain leukemias, lymphomas, and sarcomas in particular), many cancers are thought to require two or more hits [1]. For example, retinoblastoma is a two-hit cancer—indeed, a two-hit model of retinoblastoma predicted the existence of the tumor suppressing gene pRb before it was discovered [2]—and colorectal cancer can be described by three or more hits to the APC, RAS, and P53 genes [1]. Similarly to the development of precancerous polyps for colorectal cancer, many esophageal cancers begin with a transition to a condition called Barrett's esophagous [3] before accumulating additional abnormalities. These genetic (or epigenetic) hits are often described as starting different phases of carcinogenesis: initiation, the first destabilizing mutation(s); promotion, the unchecked growth of a tumor; and malignant conversion, the spread into other tissues. This classification is useful because different exposures may act on different stages of carcinogenesis.

Multistage clonal expansion (MSCE) models are a class of continuous-time Markov chain models that capture this initiation–promotion–malignant-conversion hypothesis of carcinogenesis. Originally posed as a two-stage model [4, 5] using birth–death–mutation branching-process theory, this class of models has been expanded to three or more stages, multiple pathways, and other variations. These models have been successfully used to analyze epidemiological population-level cancer incidence data [6–11], to assess the impact of time-varying exposures on cancer risk using individual-level data [12–16], and to project the impact of prevention and control strategies on population cancer rates [10, 17–19]. Although models that use multiple clonal expansion steps have been considered, models with multiple initiation stages but only a single, final clonal expansion stage are more common in the literature and appear to capture the incidence patterns of many cancers (e.g. [6–8, 20]. We are concerned here with parameter estimation for MSCE models because it can lead to better understanding of the rates of biological processes like tumor growth or adverse mutations. Indeed, knowing the approximate speed at which an abnormality arises may help to classify the

underlying abnormal event (e.g. single nucleotide mutation, chromosomal translocation, or epigenetic change).

Identifiability is the study of the parametric information available in a data set when viewed through the lens of a model, and identifiability analysis is an important precursor to accurate parameter estimation. A model is said to be *identifiable* if all model parameters may be uniquely determined from observed data [21–23]. There are two kinds of identifiability analyses: structural—which analyzes the model in the context of perfectly measured and noise-free data in order to uncover the inherent limitations of the model structure in the context of parameter estimation—and practical—which considers obstacles to parameter estimation that arise from noise, sampling frequency, bias, and other issues in real-world data sets [24]. Identifiability analysis can identify parameter combinations that embody the parametric information available in the data and lead to useful reparameterizations of the model [23].

That MSCE models are not fully identifiable is well established [6, 25–27]. In particular, finding the closed-form solution of a model's hazard function—the model output corresponding to age-specific incidence data—gives an upper bound on the number of identifiable parameter combinations available for that model from the age-specific incidence data and constrains the forms of those combinations. We previously computed the exact structural identifiability for the class of MSCE models with constant parameters and one clonal-expansion step [28]. However, this is not the last word on the identifiability of MSCE models. In particular, it is known that there is a practical identifiability problem with the clonal expansion models with three or more stages: the information contained in the asymptote of the corresponding hazard function is not available in the usual age-specific cancer incidence data because the asymptote is not reached within human lifespans [8].

In this analysis, we examine this practical identifiability problem with a profile-likelihood approach. We consider pancreatic cancer, which has linear age-specific incidence at older ages [8] and can be fit by an MSCE model with two or more stages. We demonstrate that the two-, three-, and four-stage models have only three practically identifiable parameter combinations and that, for the three- and four-stage models, several parameters that are theoretically structurally identifiable individually, are, practically, identifiable only in their product. This practical unidentifiability means the incidence data contains information about the overall rate of progression from normal to cancer-initiated cells but not the expected information on the rates of the individual steps leading to initiation.

## Methods

### Multistage clonal expansion models

The mathematics of multistage clonal expansion models have been detailed elsewhere [4, 5, 8, 25, 29–36], so we only give a basic description here. The $n$-stage clonal expansion model (Fig 1) is a continuous-time Markov chain with the following states: $X(t)$, the number of normal cells at age $t$; $Y_1(t), \ldots, Y_{n-2}$, the number of cells in subsequent preintiation states; $Y_{n-1}(t)$, the number of initiated cells; and $Z(t)$, the number of malignant cells. Let $\nu$ be the initial mutation rate, $\mu_1, \ldots, \mu_{n-3}$ the following preinitiation mutation rates, $\mu_{n-2}$ the initiation mutation rate, $\mu_{n-1}$ the malignant transformation rate, $\alpha$ the clonal expansion rate, and $\beta$ the cell death rate. If the parameters and $X(t)$ are constant, then we may denote

$$p_n, q_n := \frac{1}{2}\left(-(\alpha - \beta - \mu_{n-1}) \mp \sqrt{(\alpha - \beta - \mu_{n-1})^2 + 4\alpha\mu_{n-1}}\right), \tag{1}$$

and write hazard functions [6, 8] of the two-, three-, and four-stage models (a derivation is

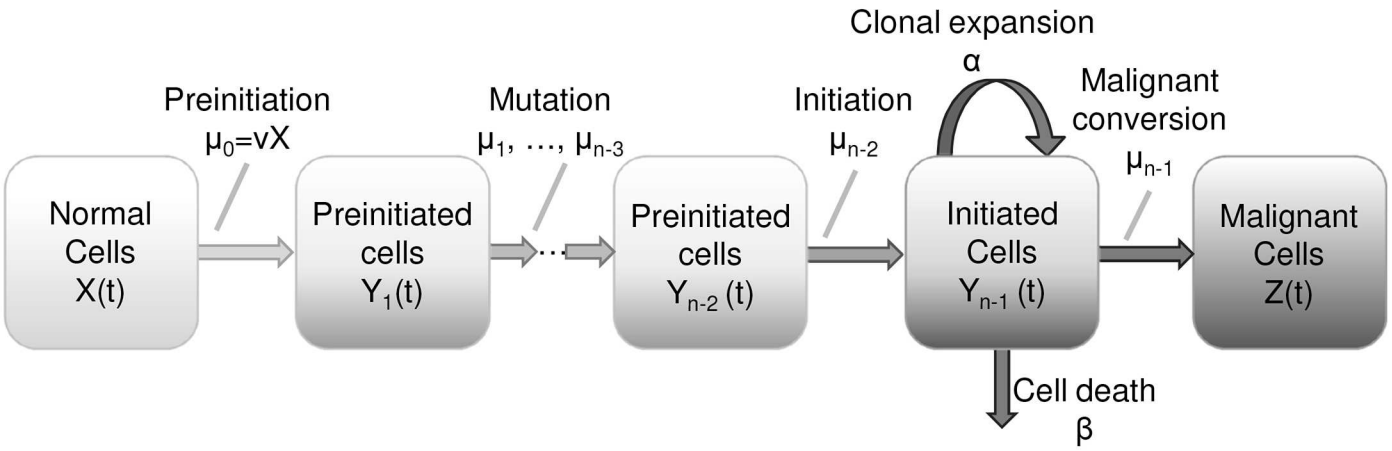


**Fig 1. Schematic of a general multistage clonal expansion model.** Multistage clonal expansion models are continuous-time Markov chain models in which normal cells undergo a series of genetic changes that lead to a state of clonal expansion followed by progression to malignancy.

https://doi.org/10.1371/journal.pcbi.1005431.g001

provided in S1 Text):

$$h_2(t) = \frac{vX}{\alpha}\left(\frac{p_2 q_2 (e^{-q_2 t} - e^{-p_2 t})}{q_2 e^{-p_2 t} - p_2 e^{-q_2 t}}\right), \tag{2}$$

$$h_3(t) = vX\left(1 - \left(\frac{q_3 - p_3}{q_3 e^{-p_3 t} - p_3 e^{-q_3 t}}\right)^{\mu_1/\alpha}\right), \tag{3}$$

$$h_4(t) = vX\left(1 - \exp\left(\int_0^t \mu_1 \left(\left(\frac{q_4 - p_4}{q_4 e^{-p_4(t-u)} - p_4 e^{-q_4(t-u)}}\right)^{\mu_2/\alpha} - 1\right) du\right)\right). \tag{4}$$

From the hazard functions, we can see that the two-, three-, and four-stage models have at most three ($vX/\alpha$, $p_2$, $q_2$), four ($vX$, $\mu_1/\alpha$, $p_3$, $q_3$), and five ($vX$, $\mu_1$, $\mu_2/\alpha$, $p_4$, $q_4$) structurally identifiable parameter combinations. In this case, these parameter combinations are structurally identifiable [28].

Multistage clonal expansion model hazards share similar characteristics, including an exponential region, a linear region, and an asymptote (Fig 2). The transition from the linear phase to the asymptote occurs on different time scales for the different models, and, for biologically reasonable ranges of the parameters, only $h_2$ can achieve this asymptote within human lifespans. The other hazards achieve their asymptotes on the order of 1,000 to 100,000 years, depending on the parameters. For example, the asymptote of the three-stage model occurs on the order of $(\mu_1(1 - \beta/\alpha))^{-1}$ [8], and mutation rate estimates are typically on the order of $10^{-7}$–$10^{-5}$ [4–8] (note that $0 < (1 - \beta/\alpha) < 1$, so that this term can only exacerbate the time span). Thus, because real data cannot access the information contained in the asymptote and other late appearing features, one may expect inherent practical identification issues for MSCE models with more stages.

## Data

We consider cancers reported to the Surveillance, Epidemiology, and End Results (SEER) cancer registries, using SEER 9 data 1973–2012 (data available in S1 and S2 Data). We use the

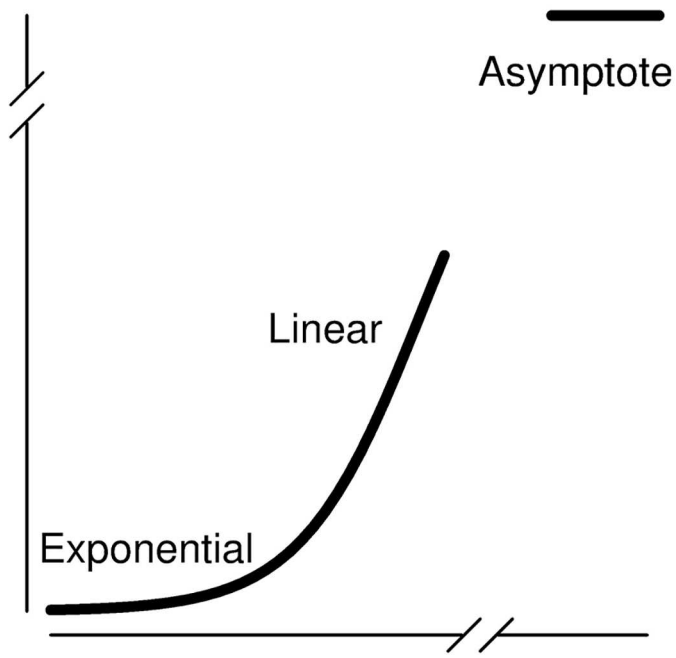


**Fig 2. Salient features of a multistage clonal expansion model hazard.** Multistage clonal expansion models have exponential and linear phases that may be observed on the time scale of a human lifespan. Depending on the number of stages in the model, the asymptote may or may not occur within a human lifespan.

https://doi.org/10.1371/journal.pcbi.1005431.g002

International Classification of Diseases (ICD-10) codes to identify incidence of pancreatic cancer (C25).

## Identifiability framework

More thorough treatments of identifiability of dynamical systems are presented elsewhere [23, 24, 37, 38], and we previously described a framework to apply dynamical systems identifiability techniques to stochastic time-to-event models, including multistage clonal expansion models [28]. Nevertheless, we provide the basic identifiability framework and definitions here for reference.

Consider a vector of states $\boldsymbol{x}(t)$ (unobserved), vector of parameters to be estimated $\boldsymbol{\rho}$, and observed (known) input $u(t)$ and output $v(t)$ in the dynamical systems model,

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= f(\boldsymbol{x}(t), u(t), \boldsymbol{\rho}), \\ v(t) &= g(\boldsymbol{x}(t), \boldsymbol{\rho}). \end{aligned} \tag{5}$$

**Definition 1** *Parameter $\rho_i$ in the model given in Eq (5) is (globally) structurally identifiable if, for almost all values $\rho_i^*$ and initial conditions, the observation of an output trajectory ($v(t) = v^*(t)$) uniquely identifies $\rho_i$ ($\rho_i = \rho_i^*$), i.e. if only one value of $\rho_i$ could have resulted in the observed output.*

**Definition 2** *The model given in Eq (5) is (globally) structurally identifiable if each $\rho_i$ is structurally identifiable.*

The definition of structural identifiability concerns perfectly measured input and output. However, because real data may not capture all of the parameteric information available in a theoretic trajectory, parameters that are structurally identifiable in a model for a kind of

theoretical data may be practically unidentifiable given a corresponding real-world dataset. Practical non-identifiability can arise from poor data quality (uncertainty, infrequent sampling, etc), but it can also be inherent to the type of data measured. For example, the saturation constant of a Michaelis-Menten equation may not be identifiable from low-dose data [39], and the amplitude of a circadian rhythm will not be identifiable if a value is measured once a day at the same time [40]. Thus, even if there are a large number of data points (e.g. as is often the case for cancer registry data), practical identifiability may still be an issue. It is this kind of inherent limitation of the data that we explore for the multistage clonal expansion models.

Practical identifiability is difficult to define in a rigorous way without choosing a threshold (e.g. width of a confidence interval) and thus has a "I know it when I see it" quality. Nevertheless, descriptions of practical identifiability are possible and typically consider the confidence bounds for the estimated parameters, found by Fisher Information Matrix (FIM) [22, 23, 41, 42] or likelihood-based methods [24]. In this analysis, we use likelihood-based confidence intervals, which are defined as follows. Let $\mathcal{L}(\boldsymbol{\rho})$ be the likelihood for the model given the data set as a function of the parameters $\boldsymbol{\rho}$, and let $\hat{\boldsymbol{\rho}}$ the maximum-likelihood estimator.

**Definition 3** *Let $\mathcal{L}^*(\rho_i)$ denote the maximum likelihood when the ith parameter is fixed to value $\rho_i$, and call it the profile likelihood of $\rho_i$. Then, the likelihood-based confidence interval for $\rho_i$ at level of significance $\alpha$ is the set of values of $\rho_i$ for which the relative negative log-likelihood at $\rho_i$ is less than a threshold determined by $\alpha$, that is,*

$$\{\rho_i : \log(\mathcal{L}(\hat{\boldsymbol{\rho}})) - \log(\mathcal{L}^*(\rho_i)) < \Delta_\alpha\}, \tag{6}$$

*where*

$$2\Delta_\alpha = \chi^2(\alpha, \mathrm{df}) \tag{7}$$

*is the chi-squared distribution with a number of degrees of freedom (df) equal to the number of parameters (for simultaneous confidence intervals) or equal to 1 (for pointwise confidence intervals).* [24, 43].

We would like to say that parameter $\rho_i$ in the model given in Eq (5) is practically identifiable if the likelihood-based confidence interval for $\rho_i$ has finite length. However, this definition is neither well-defined (the confidence interval may be finite for one level of significance but infinite at another) nor practically verifiable. Ultimately, parameters with confidence intervals that are sufficiently large—typically orders of magnitude—as to cause uncertainty and parameter estimation problems at the desired parameter scale and level of significance can be said to be *practically unidentifiable*.

## Computation methods

We use profile likelihood [24] and subset profiling [42] methods to investigate the practical identifiability of the two-, three-, and four-stage models. We assume that cancer incidence is Poisson distributed (details in S1 Text). Profile likelihoods were computed by fixing the value of one parameter at each of a series of values within an interval and numerically optimizing the negative log-likelihood as a function of the remaining parameters. Numerical optimization was done in `R` (v.3.0.1) using the `Bhat` package [44].

# Results

We plot incidence rates of pancreatic cancer reported to SEER 9 (1973–2012) in men by decade (Fig 3). The data exhibit the classic pattern of linear incidence at older ages. There are no apparent temporal trends, so we fit the two-, three-, and four-stage clonal expansion model hazards to the entire data set by minimizing the negative log-likelihood. The Akaike

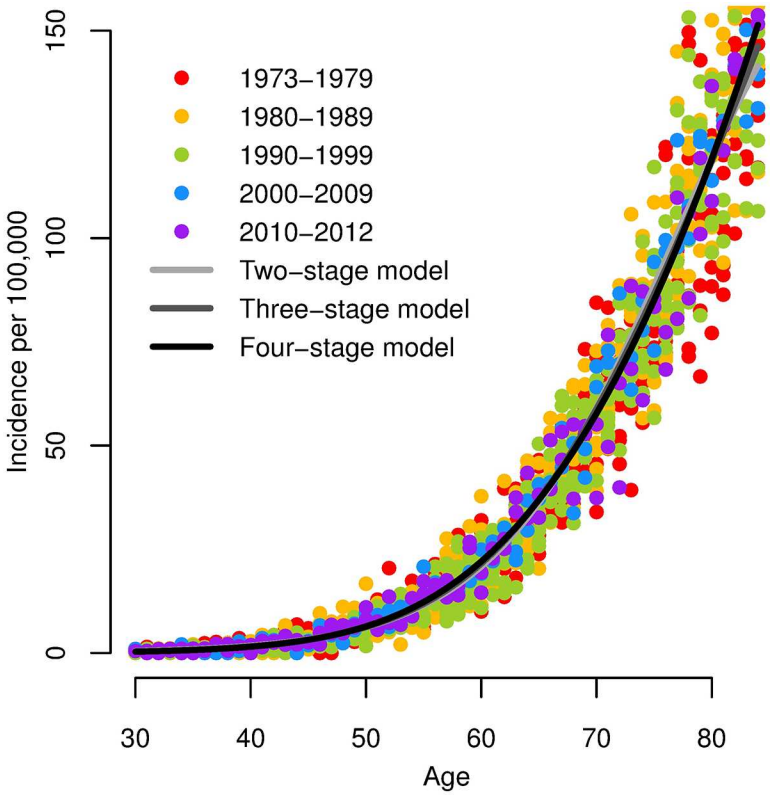


**Fig 3. Pancreatic cancer incidence and best-fit MSCE models.** Incidence of pancreatic cancer in men per 100,000 (SEER 9, 1973–2012), and best-fit two-, three-, and four-stage clonal expansion model hazards. Note that the three model hazards largely overlap.

https://doi.org/10.1371/journal.pcbi.1005431.g003

Information Criterion (AIC) for each model (relative to the best-fitting model) is 177.7, 72.3, and 0, respectively, which preferences the four-stage model.

## Two-stage model

We profile the relative negative log-likelihood of the maximum-likelihood two-stage hazard as a function of each of the parameter combinations $p_3$, $q_3$, and $\nu X/\alpha$ (Fig 4). All three parameters combinations are practically identifiable because of the trough-shape of the negative log-likelihood, giving finite confidence intervals. The parameter estimates are given in Table 1.

## Three-stage model

We profile the relative negative log-likelihood of the maximum-likelihood three-stage hazard as a function of each of the parameter combinations $p_3$, $q_3$, $\nu X$, and $\mu_1/\alpha$ (Fig 5). Parameter combinations $p_3$ and $q_3$ are practically identifiable as above, but parameter combinations $\nu X$ and $\mu_1/\alpha$ are not practically identifiable because their likelihoods flatten out, resulting in infinite confidence intervals.

To identify the form of the practically-identifiable parameter combination of $\nu X$ and $\mu_1/\alpha$, we plot the fitted value of $\mu_1/\alpha$ as we vary the value of $\nu X$ (Fig 6). Because the relationship is linear on the log–log scale, $\nu X$ and $\mu_1/\alpha$ exist in a practically identifiable product. From the biological perspective, this means that we can only know the net rate of transition from normal to initiated cells but not the rates of the individual intermediate steps.

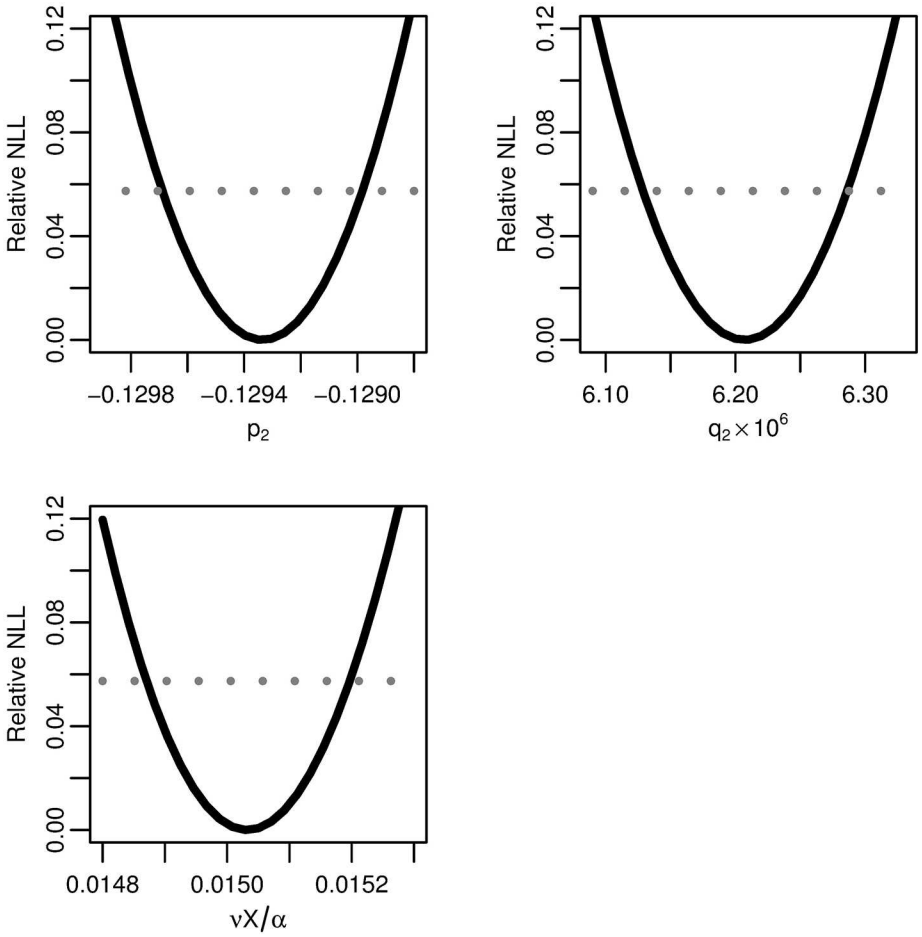


**Fig 4. Two-stage model profile likelihoods.** Profiles of the relative negative log-likelihood (NLL) of the two-stage clonal expansion model as each of the parameter combinations $p_2$, $q_2$, and $\nu X/\alpha$ are varied while the remaining parameters are fit. The gray dotted line gives the $\alpha = 0.01$ threshold for simultaneous confidence intervals based on the relative negative log-likelihood. All three parameters are identifiable.

https://doi.org/10.1371/journal.pcbi.1005431.g004

**Table 1. Best-fit parameters and likelihood-based 99% confidence intervals.** Best-fit parameters and likelihood-based 99% confidence intervals for the fits of the two-, three-, and four-stage clonal expansion models (with parameterizations using only practically identifiable parameter combinations and given in Eqs (2), (9) and (11) respectively) to age-specific incidence of pancreatic cancer.

| Model | Parameter combination | Value | Likelihood-based 99% CI |
|---|---|---|---|
| **Two-stage** | | | |
| | $p_2 = \frac{1}{2}\left(-(\alpha-\beta-\mu_1) - \sqrt{(\alpha-\beta-\mu_1)+4\alpha\mu_1}\right)$ | -1.29E-1 | (-1.30E-1, -1.29E-1) |
| | $q_2 = \frac{1}{2}\left(-(\alpha-\beta-\mu_1) + \sqrt{(\alpha-\beta-\mu_1)+4\alpha\mu_1}\right)$ | 6.21E-6 | (6.13E-6, 6.29E-6) |
| | $r_2 = \nu X/\alpha$ | 1.50E-2 | (1.49E-2, 1.52E-2) |
| **Three-stage** | | | |
| | $p_3 = \frac{1}{2}\left(-(\alpha-\beta-\mu_2) - \sqrt{(\alpha-\beta-\mu_2)+4\alpha\mu_2}\right)$ | -1.38E-1 | (-1.39E-1, -1.37E-1) |
| | $q_3 = \frac{1}{2}\left(-(\alpha-\beta-\mu_2) + \sqrt{(\alpha-\beta-\mu_2)+4\alpha\mu_2}\right)$ | 1.57E-5 | (1.53E-5, 1.60E-5) |
| | $r_3 = \sqrt{\nu X\mu_1/\alpha}$ | 2.35E-2 | (2.33E-2, 2.38E-2) |
| **Four-stage** | | | |
| | $p_4 = \frac{1}{2}\left(-(\alpha-\beta-\mu_3) - \sqrt{(\alpha-\beta-\mu_3)+4\alpha\mu_3}\right)$ | -1.50E-1 | (-1.52E-1, -1.48E-1) |
| | $q_4 = \frac{1}{2}\left(-(\alpha-\beta-\mu_3) + \sqrt{(\alpha-\beta-\mu_3)+4\alpha\mu_3}\right)$ | 4.59E-5 | (4.40E-5, 4.78E-5) |
| | $r_4 = (\nu X\mu_1\mu_2/\alpha)^{1/3}$ | 2.66E-2 | (2.63E-2, 2.70E-2) |

https://doi.org/10.1371/journal.pcbi.1005431.t001

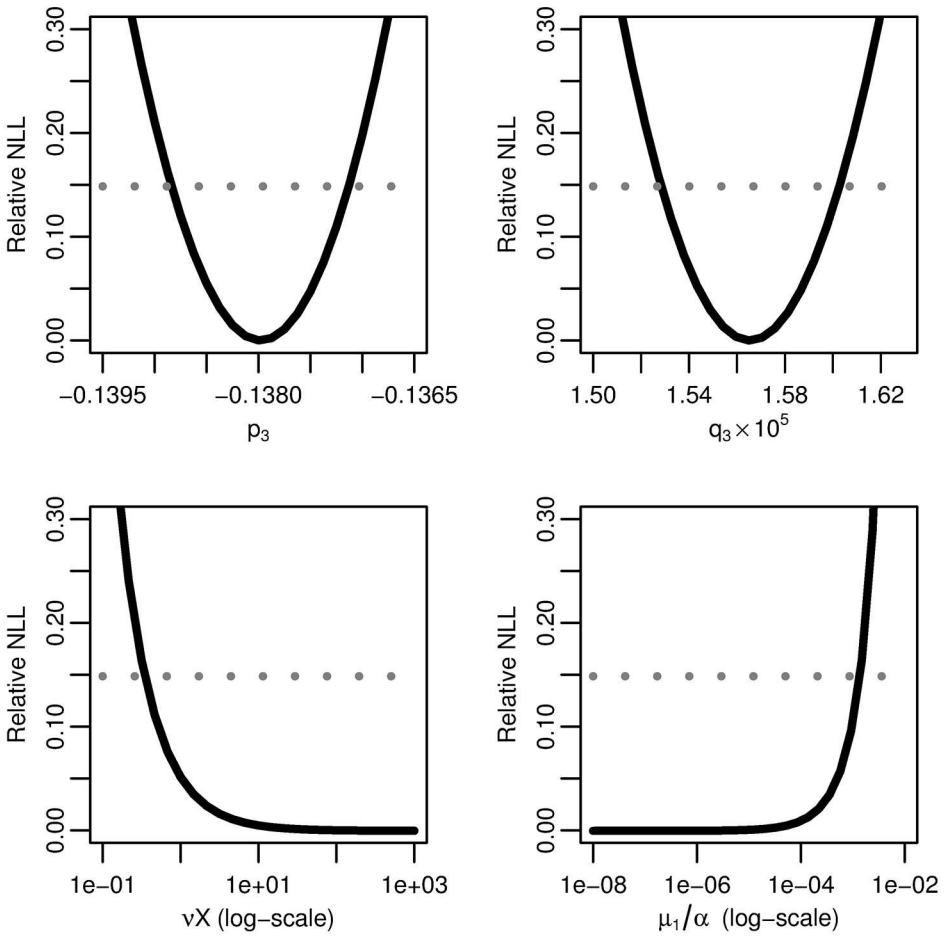


**Fig 5. Three-stage model profile likelihoods.** Profiles of the relative negative log-likelihood (NLL) of the three-stage clonal expansion model as each of parameter combinations $p_3$, $q_3$, $\nu X$, and $\mu_1/\alpha$ are varied while the remaining parameters are fit. The gray dotted line gives the $\alpha = 0.01$ threshold for simultaneous confidence intervals based on the relative negative log-likelihood. Parameter combinations $p_3$ and $p_4$ are identifiable, while $\nu X$ and $\mu_1/\alpha$ are practically unidentifiable.

https://doi.org/10.1371/journal.pcbi.1005431.g005

Our analysis thus demonstrates that there are three parameter combinations that are practically identifiable for the three-stage model from age-specific cancer-incidence data. Since there are three pieces of information in the data and four degrees of freedom in the full model (Eq 3), one might assume that one additional constraint on the model is sufficient to reduce the number of parameters estimated to three and simultaneously resolve the non-identifiability problem. However, the most reasonable simplifying assumption, namely that the first two mutation rates are the same ($\nu = \mu_1$), such as for biallelic gene inactivation [1], does not do this; the three-stage model with $\nu = \mu_1$ still has four structurally identifiable parameter combinations, namely $p_3$, $q_3$, $\nu X$, and $\nu/\alpha$, but only three pieces of practically identifiable information, so another constraint would be needed for a fully identifiable model. In this case, the constraint would need to designate the relative values of $\nu X$ and $\mu_1/\alpha$, which assuming $\nu = \mu_1$ does not do. The $\nu = \mu_1$ assumption does, however, suggest a new reparameterization of Eq 3. Denote

$$r_3 := \sqrt{(\nu X)(\mu_1/\alpha)}, \tag{8}$$

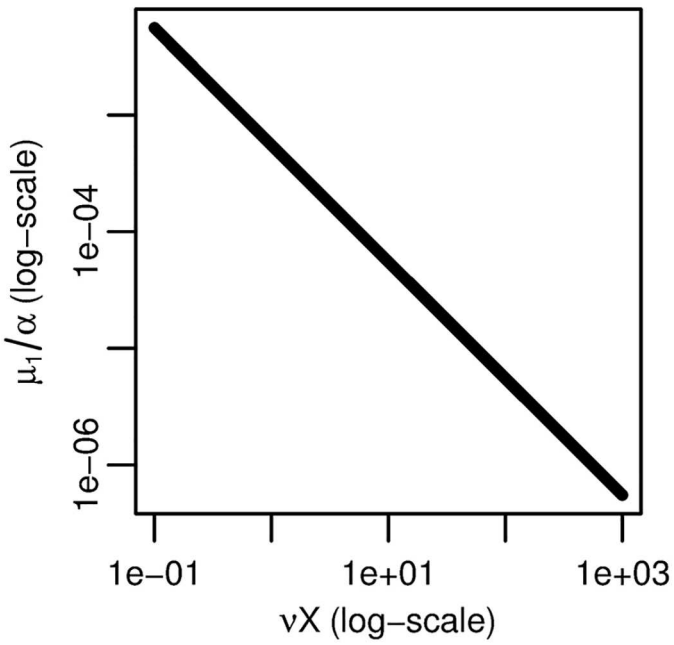


**Fig 6. Three-stage model parameter dependency.** Fitted value of $\mu_1/\alpha$ as $\nu X$ is varied for the three-stage clonal expansion model. The linear relationship on a log–log scale indicates an identifiable product.

https://doi.org/10.1371/journal.pcbi.1005431.g006

and fix $X$ and $\alpha$ at reasonable values, i.e. at values where the likelihood profiles are flat (see Fig 5). Then, assuming $\nu = \mu_1$, we parameterize $\nu X = r_3\sqrt{\alpha X}$ and $\mu_1/\alpha = r_3/\sqrt{\alpha X}$, and write

$$h_3(t) = r_3\sqrt{\alpha X}\left(1 - \left(\frac{q_3 - p_3}{q_3 e^{-p_3 t} - p_3 e^{-q_3 t}}\right)^{r_3/\sqrt{\alpha X}}\right). \quad (9)$$

As long as $X$ and $\alpha$ are chosen so that $\nu X$ and $\mu_1/\alpha$ are within a the range of values for which the likelihood is flat, their exact values do not affect the model fit and can be fixed. Caution is advisable here, however: although the exact values of these parameters do not affect the fit in this context, it is important to not take these values into other contexts where the exact values may be relevant, e.g. prediction in context of time-varying exposures. Nevertheless, this parameterization has several advantages. In particular, multiplicative effects on $r_3$, such as relative period or cohort effects, can be thought of as affecting both $\nu$ and $\mu_1$ equally: under the assumption $\mu \coloneqq \nu = \mu_1$, $r_3$ simplifies to $r_3 = \mu\sqrt{X/\alpha}$, and, more generally, we can write, for some scalar $\xi$, $\xi r_3 = \sqrt{(\xi\nu)(\xi\mu_1)(X/\alpha)}$.

We see that the profile relative NLL of $r_3 = \sqrt{\nu\mu_1 X/\alpha}$ has a finite confidence interval (Fig 7), as $p_3$ and $q_3$ did in Fig 5. The best-fit parameters for the three-stage model—parameterized as in Eq (9) and fit to the age-specific pancreatic cancer incidence data—are given in Table 1.

## Four-stage model

We similarly profile the relative negative log-likelihood of the maximum-likelihood four-stage hazard as a function of each of the parameter combinations $p_4$, $q_4$, $\nu X$, $\mu_1$, and $\mu_2/\alpha$ (Fig 8). As before, parameters combinations $p_4$ and $q_4$ have finite confidence intervals and are practically

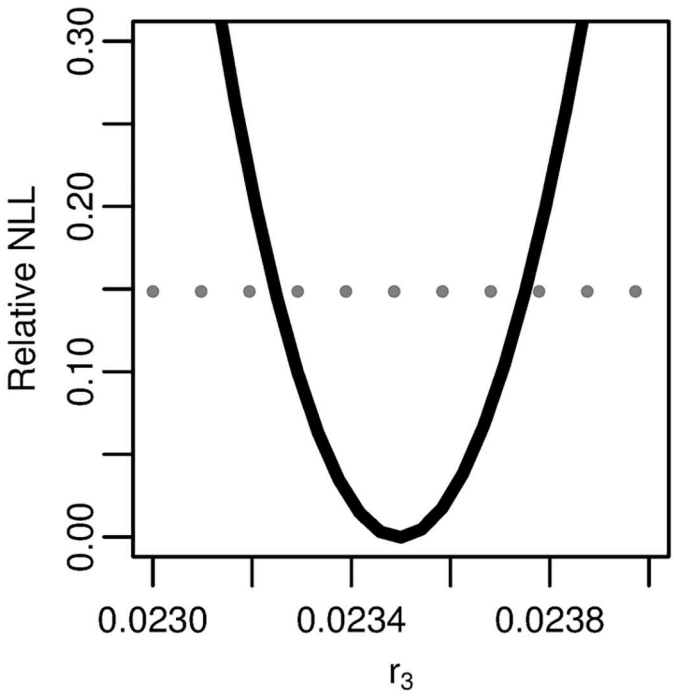


**Fig 7. Profile likelihood for the reparameterized combination of the three-stage model.** Profile of the relative negative log-likelihood (NLL) as the parameter $r_3 = \sqrt{\nu X \mu_1/\alpha}$ is varied while the remaining parameters are fit in the three-stage clonal expansion model. The gray dotted line gives the $\alpha = 0.01$ threshold for simultaneous confidence intervals based on the relative negative log-likelihood. Parameter combination $r_3$ is identifiable.

https://doi.org/10.1371/journal.pcbi.1005431.g007

identifiable, while combinations $\nu X$, $\mu_1$ and $\mu_2/\alpha$ have infinite confidence intervals and are not practically identifiable.

To determine the combination structure, we use subset profiling [42]. However, rather than using FIM to determine the profiled parameter subsets, we note that the analysis of the three stage model leads us to suspect that the three parameter combinations $\nu X$, $\mu_1$, and $\mu_2/\alpha$ are in a practical product. We use this structure to propose our nearly-full rank subsets. To verify this proposal, we plot the fitted value of one parameter combination while another is fixed and the third is varied (Fig 9). The three selected plots presented are sufficient to verify that the three parameter combinations indeed exist in a practically identifiable product. As for the three-stage case, that $\nu X$, $\mu_1$, and $\mu_2/\alpha$ can only be identified up to their product means that we can only know the net rate of transition from normal to initiated cells but not the rates of the individual intermediate steps.

We can define a quantity analogous to $r_3$ in the three stage case. Here,

$$r_4 = (\nu X \mu_1 \mu_2/\alpha)^{1/3} \tag{10}$$

and, for some reasonable fixed values of $X$ and $\alpha$,

$$h_4(t) = r_4 (X^2\alpha)^{1/3} \left(1 - \exp\left(\int_0^t r_4(\alpha/X)^{1/3} \left(\left(\frac{q_4 - p_4}{q_4 e^{-p_4(t-u)} - p_4 e^{-q_4(t-u)}}\right)^{r_4/\left(\alpha^2 X\right)^{1/3}} - 1\right) du\right)\right). \tag{11}$$

We see that the profile relative NLL of $r_4 = (\nu X \mu_1 \mu_2/\alpha)^{1/3}$ has the expected trough shape (Fig 10), as seen in Fig 8 for $p_4$ and $q_4$. The best-fit parameters for the four-stage model—

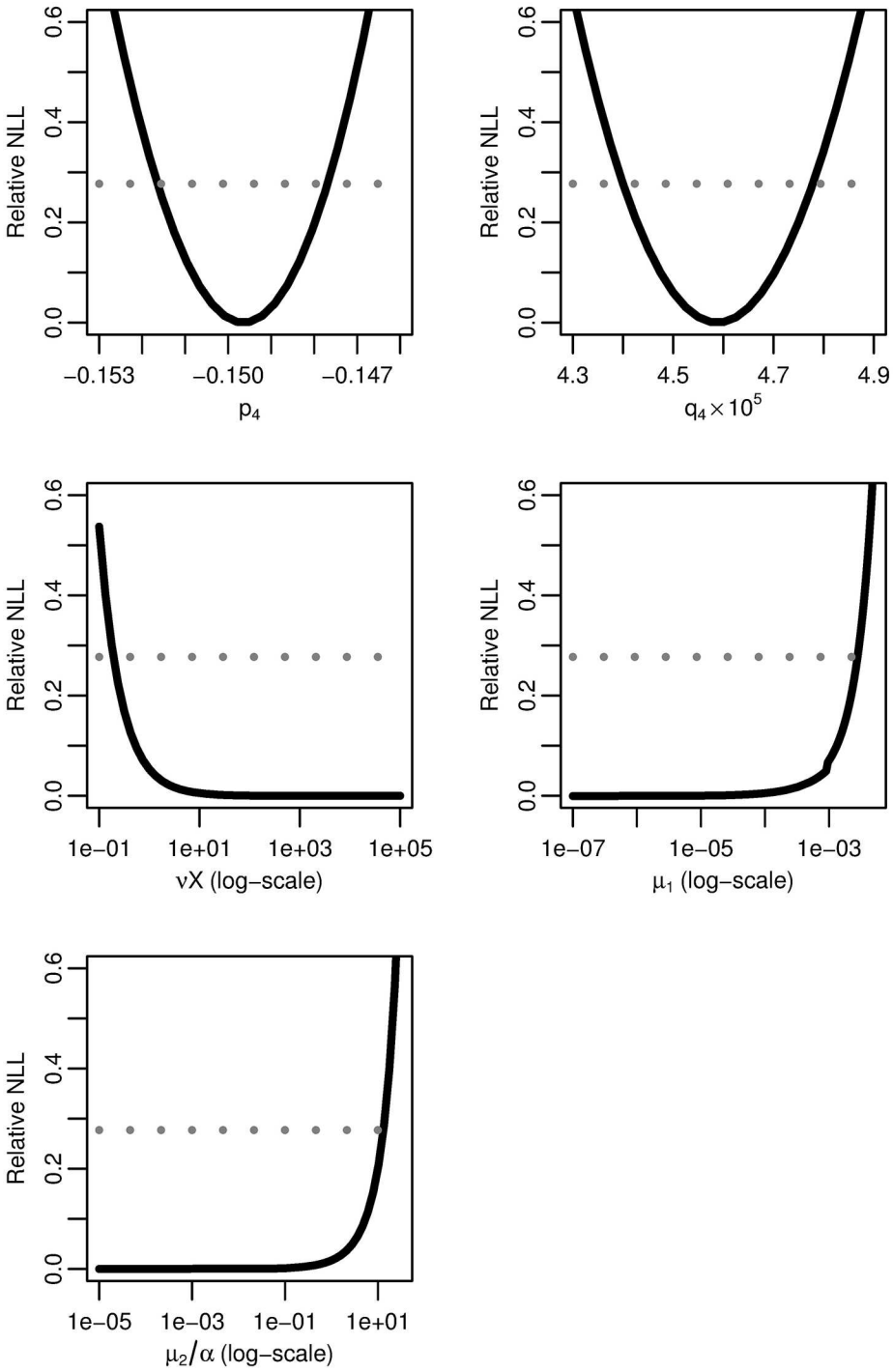


**Fig 8. Four-stage model profile likelihoods.** Profiles of the relative negative log-likelihood (NLL) of the four-stage clonal expansion model as each of parameter combinations $p_4$, $q_4$, $\nu X$, $\mu_1$, and $\mu_2/\alpha$ are varied while the remaining parameters are fit. The gray dotted line gives the $\alpha = 0.01$ threshold for simultaneous confidence intervals based on the relative negative log-likelihood. Parameter combinations $p_4$ and $q_4$ are identifiable, while $\nu X$, $\mu_1$, and $\mu_2/\alpha$ are practically unidentifiable.

https://doi.org/10.1371/journal.pcbi.1005431.g008

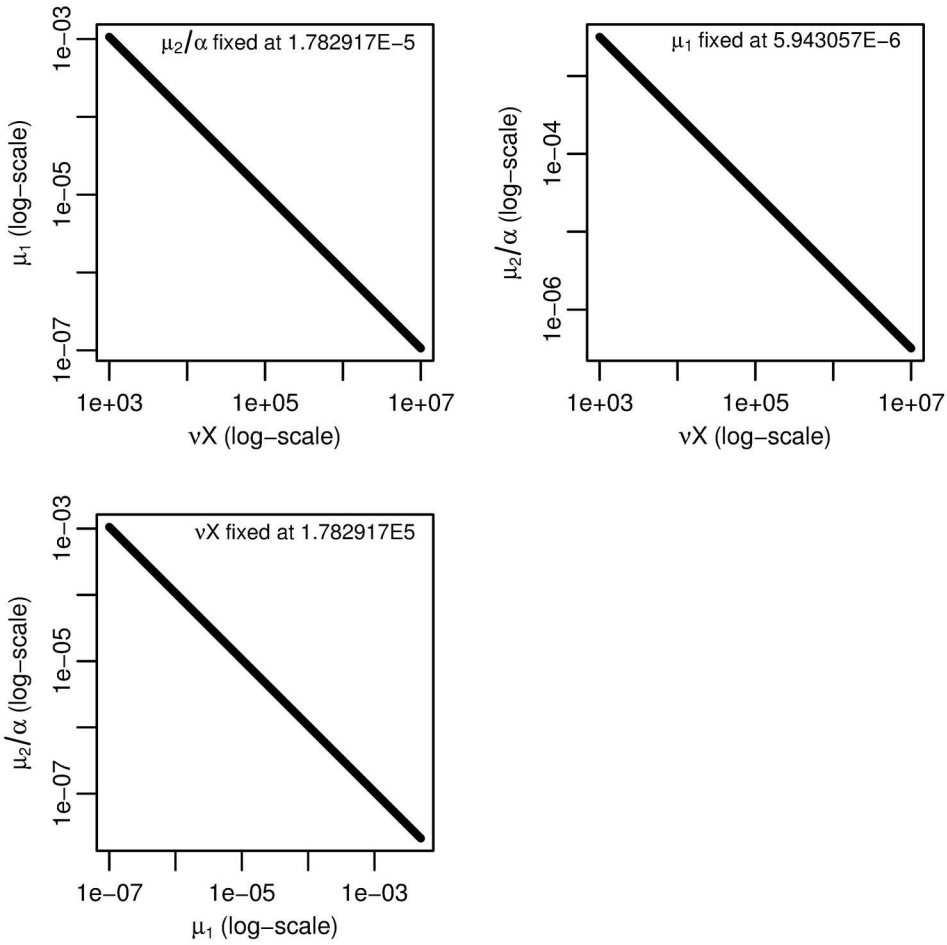


**Fig 9. Four-stage model parameter dependencies.** Fitted values of one of $\nu X$, $\mu_1$, or $\mu_2/\alpha$ as another is fixed and the third is varied for the four-stage model. The linear relationships on a log–log scale indicate that $\nu X \mu_1 \mu_2/\alpha$ is an identifiable product.

https://doi.org/10.1371/journal.pcbi.1005431.g009

parameterized as in Eq (11) and fit to the age-specific pancreatic cancer incidence data—are given in Table 1.

## Discussion

Practical unidentifiability is a significant barrier to parameter estimation. Indeed, because it—unlike structural identifiability—can be so dependent on the quality of the data, it can be a moving target. From this perspective, ironically, it is perhaps fortunate that the practical identifiability issue described herein is inherent to any age-specific cancer incidence data that is linear at older ages. This way, such problems can be anticipated and handled systematically, e.g. by reparameterizing the model appropriately. In theory, we could gain additional information if people were to live long enough to see the incidence plateau, but, as previously discussed, the expected timing of the plateau in the three- and four-stage clonal expansion models is well beyond conceivable human life spans. While the observation of a plateau might suggest that either the underlying mechanism is the two-stage model or the presence of heterogeneities or temporal effects, the absence of a plateau leaves room for various interpretations. Indeed, the

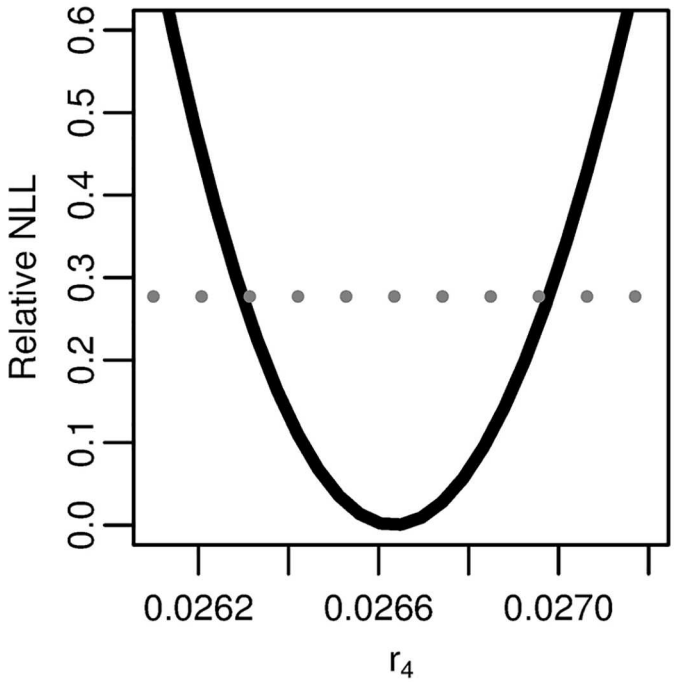


**Fig 10. Profile likelihood for the reparameterized combination of the four-stage model.** Profile of the relative negative log-likelihood (NLL) as the parameter $r_4 = (vX\mu_1\mu_2/\alpha)^{1/3}$ is varied while the remaining parameters are fit in the four stage clonal expansion model. The gray dotted line gives the $\alpha = 0.01$ threshold for simultaneous confidence intervals based on the relative negative log-likelihood. Parameter combination $r_4$ is identifiable.

https://doi.org/10.1371/journal.pcbi.1005431.g010

two-, three-, and four-stage models were all able to reasonably fit the pancreatic cancer incidence data (Fig 3).

For each of the two-, three-, and four-stage models, only three parameter combinations were practically identifiable. In each case, these combinations are most easily interpreted in the following forms:

$$\alpha - \beta - \mu_{n-1} \tag{12}$$

the net cell proliferation rate of initiated cells,

$$\alpha\mu_{n-1} \tag{13}$$

the scaled malignant conversion rate, and

$$v\left(\prod_{i=1}^{n-2}\mu_i\right)(X/\alpha) \tag{14}$$

the product of all preinitation rates scaled by the number of normal cells and the cell growth rate. Note that the first two combinations are together equivalent to $p_n$ and $q_n$. Because the last combination is a product of individually structurally identifiable combinations, we know that information about mutation rates at the intermediate steps is only available in later features of the MSCE hazards, i.e. the asymptote and the transition from the linear phase to the asymptote.

Because there are only three practically identifiable combinations, successful parameter estimation can only be achieved if the models are reparameterized in terms of these

combinations. For example, with the three-stage model parameterized as in Eq (3), parameter estimates for $vX$ and $\mu_1/\alpha$ are not stable. Here, we proposed one possible solution with the reparameterizations in Eqs (9) and (11) and show that it does indeed resolve the practical unidentifiabilities, though an infinite number of reparameterizations will give equivalent fits as long as the parameter combinations are preserved. Each reparameterization represents a different assumption about the relative sizes of its constituent parameter combinations. Our reparameterizations are inspired by the assumption that the preinitation mutation rates are equal ($v = \mu_1 = \ldots$) but do not actually codify this assumption in the models. Nevertheless, it is consistent with a scenario in which multiple copies of a tumor suppressor gene must be "knocked out" [1].

Traditional approaches to parameter estimation that use asymptotic confidence intervals do not always reveal practical identifiability issues. Because asymptotic confidence intervals are based on the local curvature of the likelihood around the parameter estimate, they may give finite confidence bounds when the likelihood is curved on one side of the estimate but flat on the other. Numerical optimization algorithms may provide results that give the appearance of practical identifiability but have in fact simply pushed the estimate to the point where the likelihood begins to curve. Hence, the fact that our group and others have previously reported values of $\mu_2/\alpha$ with finite confidence intervals in four-stage models [6, 8] is not inconsistent with our results. Some of these previous works have interpreted the larger-than-expected values for $\mu_2$ (fixing $\alpha$) as being too fast to represent a genetic mutation, suggesting that the four-stage model may represent two, slow genetic mutations followed by a fast epigenetic change, a transient event, or other transformation. Our results suggest that a large range of values $\mu_2/\alpha$ would have resulted in equivalent fits, and we note that the values presented in these previous works are of the same order of magnitude where we see curvature in our likelihood function. In particular, a previous fit of pancreatic cancer incidence in SEER (1973–2004) using the four-stage model [8] estimated $vX\mu_1\mu_2/\alpha$ to be 1.88E-5—the same value that we find here with the new parametrization (for pancreatic cancer in SEER 1973–2012; Table 1)—but also separately estimated $\mu_2/\alpha$ to be 4.0E-1, which falls exactly where the profile likelihood begins to curve up (Fig 8). Hence, such parameter estimates may be an artifact of the algorithm numerically optimizing the likelihood, and one should then be careful when giving a biological interpretation to those results.

This analysis also speaks to the question of model selection and model reduction. Although the four-stage model gives the best statistical fit to the data in Fig 3, its hazard nearly entirely overlaps with the that of the other models. Hence, we must question whether or not the larger model is actually capturing some nuance in the data. Given the practical identifiability issues we have presented, does the two-stage already capture all of the information? Possibly so. Are the results of both models equivalent? Unfortunately not: although each model is estimating the same biological parameters (i.e. the product of initiation rates, the final promotion rate, and the malignant conversion rate), a perusal of Table 1 reveals that the parameter estimates are not particularly consistent across the three models (although are generally within an order of magnitude). Moreover, the different dynamics of each model will become important as we move away from simply analyzing incidence and consider prediction or individual time-varying exposures. Nevertheless, in this situation, one might be inclined to take an ensemble approach and to consider uncertainty quantification not only within a model but across the models, perhaps weighting in some way by statistical fit. Additional empirical science, by better elucidating carcinogenesis mechanisms common to cancer at given site, could aid modelers in model selection.

The guidance we have presented in this study is important as three- and four-stage clonal expansion models are commonly used to model certain cancers at the population level, and

successful parameter estimation is dependent on the model being identifiable with respect to the available data. Ultimately, our analysis demonstrates the need for future studies to verify the practical identifiability of model parameters whenever feasible, which should strengthen the validity of the analyses and aid in the interpretation of estimated parameter values and modeling results.

## Supporting information

**S1 Data. Incidence.** Cases of pancreatic cancer in men reported to SEER 9, 1973–2012.
(CSV)

**S2 Data. Population.** Population of men in SEER 9 catchment, 1973–2012.
(CSV)

**S1 Text. Mathematical and statistical details.** Derivation of multistage clonal expansion model hazards and statistical formulation of the likelihood.
(PDF)

## Author Contributions

**Conceptualization:** AFB MCE RM.

**Funding acquisition:** MCE RM.

**Investigation:** AFB.

**Methodology:** AFB MCE RM.

**Software:** AFB.

**Validation:** AFB MCE RM.

**Writing – original draft:** AFB.

**Writing – review & editing:** AFB MCE RM.

## References

1. Knudson AG. Two genetic hits (more or less) to cancer. Nature reviews Cancer. 2001; 1(2):157–162. https://doi.org/10.1038/35101031 PMID: 11905807
2. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences. 1971; 68(4):820–3. https://doi.org/10.1073/pnas.68.4.820
3. Wild C, Hardie L. Reflux, Barrett's oesophagus and adenocarcinoma: burning questions Nature Reviews. Cancer. 2003; 3(9):676–84. PMID: 12951586
4. Moolgavkar SH, Venzon DJ. Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. Mathematical Biosciences. 1979; 47(1–2):55–77. https://doi.org/10.1016/0025-5564(79)90005-1
5. Moolgavkar SH, Knudson AG. Mutation and cancer: a model for human carcinogenesis. Journal of the National Cancer Institute. 1981; 66(6):1037–52. https://doi.org/10.1093/jnci/66.6.1037 PMID: 6941039
6. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. Proceedings of the National Academy of Sciences. 2002; 99(23):15095–15100. https://doi.org/10.1073/pnas.222118199
7. Jeon J, Luebeck EG, Moolgavkar SH. Age effects and temporal trends in adenocarcinoma of the esophagus and gastric cardia (United States). Cancer Causes & Control. 2006; 17(7):971–81. https://doi.org/10.1007/s10552-006-0037-3
8. Meza R, Jeon J, Moolgavkar SH, Luebeck EG. Age-specific incidence of cancer: Phases, transitions, and biological implications. Proceedings of the National Academy of Sciences. 2008; 105(42):16284–9. https://doi.org/10.1073/pnas.0801151105

9. Hazelton WD, Curtius K, Inadomi JM, Vaughan TL, Meza R, Rubenstein JH, et al. The role of gastro-esophageal reflux and other factors during progression to esophageal adenocarcinoma. Cancer Epidemiology, Biomarkers & Prevention 2015; 24(7):1–6.
10. Curtius K, Hazelton WD, Jeon J, Luebeck EG. A Multiscale Model Evaluates Screening for Neoplasia in Barrett's Esophagus. PLOS Computational Biology. 2015; 11(5):e1004272. https://doi.org/10.1371/journal.pcbi.1004272 PMID: 26001209
11. Brouwer AF, Meza R, Eisenberg MC. Age Effects and Temporal Trends in HPV-Related and HPV-Unrelated Oral Cancer in the United States: A Multistage Carcinogenesis Modeling Analysis. PLOS One. 2016; 11(3): e0151098. https://doi.org/10.1371/journal.pone.0151098 PMID: 26963717
12. Luebeck EG, Heidenreich WF, Hazelton WD, Paretzke HG, Moolgavkar SH. Biologically based analysis of the data for the Colorado uranium miners cohort: age, dose and dose-rate effects. Radiation Research. 1999; 152(4):339–51. https://doi.org/10.2307/3580219 PMID: 10477911
13. 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. Radiation Research. 2001; 156(1):78–94. https://doi.org/10.1667/0033-7587(2001)156%5B0078:AOAHCO%5D2.0.CO;2 PMID: 11418076
14. Meza R, Hazelton WD, Colditz GA, Moolgavkar SH. Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage carcinogenesis model. Cancer Causes & Control. 2008; 19(3):317–28. https://doi.org/10.1007/s10552-007-9094-5
15. Schöllnberger H, Beerenwinkel N, Hoogenveen R, Vineis P. Cell selection as driving force in lung and colon carcinogenesis. Cancer Research. 2010; 70(17):6797–803. https://doi.org/10.1158/0008-5472.CAN-09-4392 PMID: 20656803
16. Richardson DB. Multistage modeling of leukemia in benzene workers: a simple approach to fitting the 2-stage clonal expansion model. American Journal of Epidemiology. 2009; 169(1):78–85. https://doi.org/10.1093/aje/kwn284 PMID: 18996834
17. 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. Journal of the National Cancer Institute. 2012; 104(7):541–548. https://doi.org/10.1093/jnci/djs136 PMID: 22423009
18. 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. Annals of Internal Medicine. 2014; 160(5):311–320. https://doi.org/10.7326/M13-2316 PMID: 24379002
19. Kong CY, Kroep S, Curtius K, Hazelton WD, Jeon J, Meza R, et al. Exploring the Recent Trend in Esophageal Adenocarcinoma Incidence and Mortality Using Comparative Simulation Modeling. Cancer Epidemiology, Biomarkers & Prevention. 2014; 23(6):997–1006. https://doi.org/10.1158/1055-9965.EPI-13-1233
20. 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 Research. 2010; 70(13):5419–29. https://doi.org/10.1158/0008-5472.CAN-09-4417 PMID: 20530677
21. Bellman R, Åström KJ. On structural identifiability. Mathematical Biosciences. 1970; 7(3):329–339. https://doi.org/10.1016/0025-5564(70)90132-X
22. Rothenberg TJ. Identification in Parametric Models. Econometrica. 1971; 39(3):577–591. https://doi.org/10.2307/1913267
23. Cobelli C, DiStefano JJ. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. American Journal of Physiology. 1980; 239(1):R7–R24. PMID: 7396041
24. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009; 25(15):1923–1929. https://doi.org/10.1093/bioinformatics/btp358 PMID: 19505944
25. Heidenreich WF, Luebeck EG, Moolgavkar SH. Some properties of the hazard function of the two-mutation clonal expansion model. Risk Analysis. 1997; 17(3):391–9. https://doi.org/10.1111/j.1539-6924.1997.tb00878.x PMID: 9232020
26. Cox LA, Huber WA. Symmetry, identifiability, and prediction uncertainties in multistage clonal expansion (MSCE) models of carcinogenesis. Risk Analysis. 2007; 27(6):1441–1453. https://doi.org/10.1111/j.1539-6924.2007.00980.x PMID: 18093045
27. Little MP, Heidenreich WF, Li G. Parameter identifiability and redundancy in a general class of stochastic carcinogenesis models. PLOS One. 2009; 4(12):1–6. https://doi.org/10.1371/journal.pone.0008520
28. Brouwer AF, Meza R, Eisenberg MC. A Systematic Approach to Determining the Identifiability of Multistage Carcinogenesis Models. Risk Analysis. 2016. https://doi.org/10.1111/risa.12684

29. Dewanji A, Venzon DJ, Moolgavkar SH. A stochastic two-stage model for cancer risk assessment. II. The number and size of premalignant clones. Risk analysis: an official publication of the Society for Risk Analysis. 1989; 9(2):179–187. https://doi.org/10.1111/j.1539-6924.1989.tb01238.x

30. Moolgavkar S, Luebeck G. Two-event model for carcinogenesis: Biological, mathematical, and statistical considerations. Risk Analysis. 1990; 10(2):323–341. https://doi.org/10.1111/j.1539-6924.1990.tb01053.x PMID: 2195604

31. Tan WY. Stochastic Models of Carcinogenesis. New York: Marcel Dekker; 1991.

32. Little MP. Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon, and Knudson, and of the Multistage Model of Armitage and Doll. Biometrics. 1995; 51(4):1278–1291. https://doi.org/10.2307/2533259 PMID: 8589222

33. Heidenreich WF. On the parameters of the clonal expansion model. Radiation and Environmental Biophysics. 1996; 35(2):127–129. https://doi.org/10.1007/BF02434036 PMID: 8792461

34. Crump KS, Subramaniam RP, Van Landingham CB. A numerical solution to the nonhomogeneous two-stage MVK model of cancer. Risk Analysis. 2005; 25(4):921–6. https://doi.org/10.1111/j.1539-6924.2005.00651.x PMID: 16268939

35. Meza R. Some Extensions and Applications of Multistage Carcinogenesis Models. University of Washington; 2006.

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

37. Audoly S, Bellu G, D'Angiò L, Saccomani MP, Cobelli C. Global identifiability of nonlinear models of biological systems. IEEE Transactions on Biomedical Engineering. 2001; 48(1):55–65. https://doi.org/10.1109/10.900248 PMID: 11235592

38. Saccomani MP, Audoly S, Bellu G, D'Angio L. A new differential algebra algorithm to test identifiability of nonlinear systems with given initial conditions. Proceedings of the 40th IEEE Conference on Decision and Control. 2001;4:3108–3113.

39. Holmberg A. On the practical identifiability of microbial growth models incorporating Michaelis–Menten type nonlinearities. Mathematical Biosciences. 1982; 62(1):23–43. https://doi.org/10.1016/0025-5564(82)90061-X

40. Eisenberg M, Samuels M, DiStefano JJ. Extensions, Validation, and Clinical Applications of a Feedback Control System Simulator of the Hypothalamo-Pituitary-Thyroid Axis. Thyroid. 2008; 18(10):1071–1085. https://doi.org/10.1089/thy.2007.0388 PMID: 18844475

41. Cintrón-Arias A, Banks HT, Capaldi A, Lloyd AL. A Sensitivity Matrix Based Methodology for Inverse Problem Formulation. Journal of Inverse and Ill-posed Problems. 2009; 17(6):545–565.

42. Eisenberg MC, Hayashi MAL. Determining identifiable parameter combinations using subset profiling. Mathematical Biosciences. 2014; 256:116–126. https://doi.org/10.1016/j.mbs.2014.08.008 PMID: 25173434

43. Keener RW. Theoretical Statistics. Springer Texts in Statistics. New York, NY: Springer New York; 2010.

44. Luebeck G, Meza R. Bhat: General likelihood exploration; 2013. R package version 0.9–10. Available from: http://CRAN.R-project.org/package=Bhat.