Theoretical Biology and Medical Modelling

BioMed Central

Research

Open Access

Heterogeneity in multistage carcinogenesis and mixture modeling

Sandro Gsteiger* and Stephan Morgenthaler

Address: Institute of Mathematics, Swiss Federal Institute of Technology, Lausanne, Switzerland

Email: Sandro Gsteiger* - sandro.gsteiger@a3.epfl.ch; Stephan Morgenthaler - stephan.morgenthaler@epfl.ch

* Corresponding author

Published: 21 July 2008

Theoretical Biology and Medical Modelling 2008, **5**:13 doi:10.1186/1742-4682-5-13

Received: 26 October 2006
Accepted: 21 July 2008

This article is available from: http://www.tbiomed.com/content/5/1/13

Abstract

Carcinogenesis is commonly described as a multistage process, in which stem cells are transformed into cancer cells via a series of mutations. In this article, we consider extensions of the multistage carcinogenesis model by mixture modeling. This approach allows us to describe population heterogeneity in a biologically meaningful way. We focus on finite mixture models, for which we prove identifiability. These models are applied to human lung cancer data from several birth cohorts. Maximum likelihood estimation does not perform well in this application due to the heavy censoring in our data. We thus use analytic graduation instead. Very good fits are achieved for models that combine a small high risk group with a large group that is quasi immune.

Introduction

Cancers can arise in virtually any part of the body, and although there are many tissue specific properties, a general multistage framework for carcinogenesis holds for most cancer types. More precisely, cells must undergo an evolutionary process involving several stages and leading finally to a cell that has completely lost proliferation control. In a first step, called initiation, mutations transform stem cells into intermediate states. Such initiated cells may give rise to pre-neoplastic lesions via accelerated growth. Eventually, a cell out of such a clone may experience further mutations and be transformed into a malignant tumor cell. This second step comprising clonal expansion and final malignant transformation is commonly called promotion. This multistage scheme shows the inherently random aspect of carcinogenesis: mutations happen at random times and stochastic growth processes are involved.

Mathematical models of carcinogenesis have been studied for about fifty years. Some of the earliest attempts to build biologically based quantitative descriptions are [1] and [2], who explained cancer as the result of a sequence of mutations. A widely accepted model was proposed in [3] and [4]. Their two-stage clonal expansion (TSCE) model was explicitly formulated in terms of an initiation stage and a promotion stage. This approach stressed the importance of both mutations and clonal expansion in the process leading to cancer. The TSCE model has found many applications and extensions. One example is the multistage model, which takes up the same structure but allows for more than two stages. Due to this long and evolving story, we should not have in mind a single model when talking about the multistage model. We should rather have in mind a cascade of nested models that starts from a fundamental idea and incorporates through its evolution more and more biological detail. Excellent reviews of stochastic carcinogenesis modeling can be found in [5] and [6].

One part of the recent extensions tries to take population heterogeneity into account. Such heterogeneity can result from sources such as genetic variation, exposure to carcinogens due to either changes in environment or occupation, and differences in lifestyle (the most prominent factors being smoking and diet). In [7] a mixture of a one stage model and a two stage model was used to describe the heritable and the sporadic form of Retinoblastoma, a cancer of the eye caused by mutations in a single tumor suppressor gene. Other approaches incorporate heterogeneity via standard frailty modeling, where the common baseline hazard $h_0(t)$ is multiplied by a non-negative random variable Z in order to model the individual hazard $h_{ind}(t) = Zh_0(t)$, see for example [8], and [9] for such an approach.

In this text, we take up the work by [10]. These authors introduce two new population parameters to describe heterogeneity. The first one, called the fraction at risk F, is used to distinguish between susceptibles and a postulated group of immune individuals. The second one, called the fraction of deaths due to cancer among all deaths due to either cancer or related competing causes f, models competing related risks. They fit their model to US lung cancer incidence data from several birth cohorts. The parameters F and f present an abstract way describe population heterogeneity and are not linked to a specific biological process. Therefore, the above mentioned authors state that other modeling strategies could be tested. The present work gives such an attempt. We take up the same multistage model, but we will use mixture models to allow for variability among individuals. This allows us to introduce heterogeneity in a biologically meaningful way.

In the next section, we describe the multistage carcinogenesis model and introduce an extension by mixture. Then we will give a series of identifiability results for both the multistage model and some mixture models. Finally, we apply the model to human incidence data before giving some concluding remarks.

Mathematical Model Formulation

The Multistage Carcinogenesis Model

We will work with a simplified version of the multistage model, but one that is general enough to incorporate the two main features of the carcinogenesis process: the sequence of mutations and the clonal expansion. We make the following assumptions:

1. A cell must undergo n mutational events to get initiated.

2. The number of cells at risk, N_0, is constant over time.

3. The number of newly generated initiated cells is a (non-homogeneous) Poisson process with intensity $\lambda_I(t)$.

4. An initiated cell gives rise to a clonal expansion according to a birth-and-death process with emigration, i.e. in a short time interval $(t, t + \Delta t)$ an initiated cell divides in two initiated cells with rate β, dies or differentiates with rate $\delta(<\beta)$, and divides into one initiated and one malignant cell with rate μ.

5. Once a promoted cell is generated, its growth is deterministic, and we neglect the time needed to grow to detectable tumor size.

6. The system starts with all at risk cells in the normal state and the different cells act independently of one another.

The model is shown schematically in Figure 1. Note that the above assumptions are standard in carcinogenesis modeling, and the possibility of generalization (for example to time dependent N_0) has been discussed by several authors. However, we choose this simplified version in order to limit the complexity of our baseline model. Note also that for $n = 1$ we get the classical TSCE model.

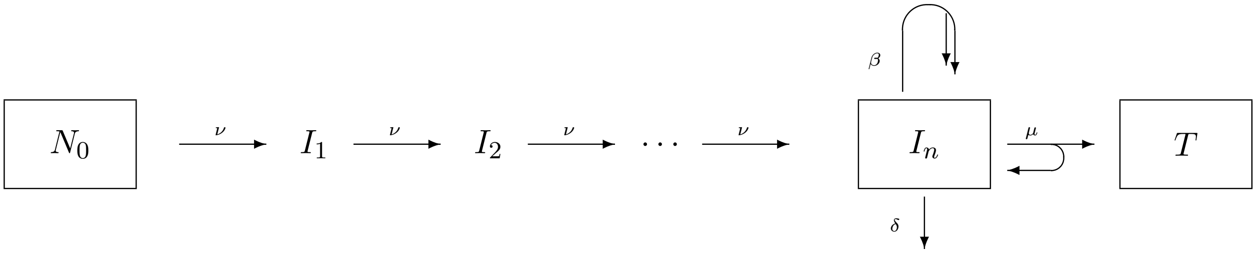

Figure 1
The multistage carcinogenesis model. N_0 denotes the number of normal stem cells. To get initiated, a normal cell accumulates n consecutive mutations, where ν denotes the mutation rate per cell per year for the gene in question. The number of cells having k mutations is noted I_k, $1 \le k \le n$. The fully initiated cells, I_n, expand according to a birth-and-death process. These cells give rise to tumor cells T if a further event happens, and $\mu/(\mu + \beta)$ can be interpreted as the probability of such a malignant transformation during a cell division.

A detailed discussion and derivation of the survivor and hazard functions of this multistage model can be found in [11,12] and [10]. These authors show that the survivor function for tumor onset can be represented as

$$S(t;n) = \exp\left\{-\int_0^t \lambda_I(x)F_P(t-x)\mathrm{d}x\right\}. \qquad (1)$$

In this expression

$$\lambda_I(x) = n\,\nu^n N_0 x^{n-1} \qquad (2)$$

is the intensity of initiation. The function $F_P(x)$ is the cdf for the waiting time for the first malignant transformation within a clone starting with one initiated cell at time 0. This cdf is improper since a clone of initiated cells dies out with a probability greater than 0 if $\delta > 0$. Its exact form is

$$F_P(x) = \frac{(\theta+\Delta)(\theta-\Delta)(e^{-\Delta x}-1)}{2\beta[(\theta+\Delta)e^{-\Delta x}-(\theta-\Delta)]}, \qquad (3)$$

where $\theta = \beta - \delta - \mu$ and $\Delta = \sqrt{(\beta+\delta+\mu)^2 - 4\beta\delta}$.

Researchers have expressed concern about the approximations used in carcinogenesis models. This issue was raised in a review paper by [13] and has inspired [11] and [14]. Our formula above is exact and uses the method based on integration cited by [[11], top of p. 1080]. The remaining simplifications in our model, in particular the constancy of N, are for convenience and do not affect the conclusions of the paper. As a general comment, it should be noted that the term two-stage refers to different things in different papers. It is, for example, possible to model clones via compartments or via branching processes. Both may employ the same parameter notation, but the interpretation will be quite different. Care has to be taken, if one wishes to stay close to biological reality. For further comments, see [[13], section 2.1].

The hazard function can be easily calculated from the survivor function, $h(t;\, n) = -\mathrm{d}\log S(t;\, n)/\mathrm{d}t$. In order to deduce the asymptotic behavior of the hazard for several n, we note that $h(t;\, n)$ can be written in terms of a recursion, $h(t;\, 1) = \nu N_0 F_P(t)$ and $h(t;\, n) = n\nu\int_0^t h(u;n-1)\mathrm{d}u$ for $n \geq 2$. Therefore, when $n = 1$ the hazard levels off as t goes to infinity. More precisely, the hazard of the TSCE model goes to the finite asymptote

$$\nu N_0 \cdot \mathrm{P}(\text{a clone of initiated cells does not die out}).$$

But the hazard grows to infinity if $n \geq 2$. In both cases $h(t;\, n)$ is strictly monotonic increasing with t. The unboundedness of the hazard is due to the simplifications in the model. This may lead to appreciable differences at low ages in some types of cancer or under hightened exposure. It is typically of lesser importance in human studies.

The monotonicity properties of hazard curves are not in agreement with observed incidence curves from human population data. Such data typically shows very low incidence up to about the age of fifty, a sudden and sharp increase between about the ages of fifty and eighty, and a subsequent leveling off and a decrease for the very old. This behavior at old ages is not captured by the hazard curves $h(t;\, n)$. However, it can be modeled very easily by incorporating a frailty effect as we will show in the application later on.

Extension of the Model by the Use of Mixture Distributions

An observed human population is heterogeneous. Though the process of cancer development is similar for everyone, parameters may vary between individuals. Since all parameters of the model are biologically meaningful, we aim at modelling heterogeneity directly through these parameters. We thus propose to consider some of the biological parameters – one at a time – as random variables.

Let θ be such a parameter and let $G(\theta)$ be a distribution function for θ. Then, we will denote by $S(t|\theta)$ the survivor function of the multistage model (1) for a given value θ, whereas the population survivor function is

$$S(t) = \int S(t|\theta)\mathrm{d}G(\theta). \qquad (4)$$

Under certain regularity conditions an analogous representation holds for the hazard function

$$h(t) = \int h(t\mid\theta)\frac{S(t|\theta)}{S(t)}\mathrm{d}G(\theta).$$

The distribution function G must then be selected based on the biological parameter θ chosen. If we consider for example $\theta = n$, the number of mutations needed for initiation, it is natural to choose a finite distribution, i.e. $\mathrm{P}(\theta = \mathrm{n_i}) = \pi_\mathrm{i}$ for a fixed set $\{n_1,...,n_g\} \subset$ such that $\sum \pi_i = 1$. This would correspond to g population subgroups having inherited different numbers of initiating mutations. The model could also be interpreted as a multiple pathway model, where g pathways involving different numbers of mutations can lead to cancer. Other interesting choices focus on promotion, for example $\theta = \beta - \delta$, the growth advantage of initiated cells, or $\theta = \mu$, the promotion rate.

These two cases would be consistent with both a finite or a continuous distribution as long as its support is contained within biologically reasonable bounds.

Identifiability

Before fitting our mixture model to observed incidence data, the identifyability issue has to be considered. The parameters of the TSCE model cannot be uniquely determined based on incidence data. Some of the papers relevant to this issue are [15,16], and [17].

Identifiability of the Multistage Model

The parameters of our base model (Eq. 1, 2, 3) are n, the number of initiating mutations, and $\psi = (N_0, \nu, \beta, \delta, \mu)$, sizes and rates. When fitting $S(t|n, \psi)$, the survivor function of the multistage model given in (1), these six parameters cannot all be fitted separately. We will show that n and the follwing three combinations are, however, uniquely determined

$$\begin{cases} p &= \beta/(\nu^n N_0), \\ q &= \delta - \beta + \mu, \\ r &= (\beta + \delta + \mu)^2 - 4\beta\delta. \end{cases}$$

In order to deal with the discrepancy between the six parameters and the four identifyable features, we will hold the two parameters N_0 and δ fixed (see also [18]). To determine N_0 = *Number of stem cells in a given tissue* some preliminary biological estimate is needed. For the death rate we mainly focus on the choice $\delta = 0$, which implies $\gamma = \beta - \delta = \beta$, so that β simply describes the growth advantage of initiated cells.

The Number of Mutations for Initiation

Is n identifiable in the multistage model or could a change in n be compensated by some adjustment of the biological parameters ψ so that finally the same survivor function resulted? The answer to this question is no, because the behavior of $S(t|n, \psi)$ at the origin is enough to determine n.

Proposition 1. *If for two parameter choices* (n, ψ), *and* $(\tilde{n}, \tilde{\psi})$ *we have* $S(t|n, \psi) \equiv_t S(t|\tilde{n}, \tilde{\psi})$, *then* $n = \tilde{n}$.

Proof: Direct calculation shows that

$$\begin{aligned} S^{(k)}(0 \mid n, \psi) &= 0 \quad \text{for } k = 1, 2, \ldots, n, \text{ and} \\ S^{(n+1)}(0 \mid n, \psi) &\neq 0, \end{aligned}$$

for all $n \in \{1, 2,...\}$, where $S^{(k)} = d^kS/dt^k$. The proposition is a direct consequence.

This result is mainly of theoretical interest and cannot be used to estimate n. In practice, for some tissues one can fix n according to the available biological theory. For example in colon cancer it is commonly assumed that two mutations are necessary for initiation, followed by a third one for malignant transformation (see [19]). In cases where no biological reasoning is available, we suggest to fit the model for several choices of n. The form of the intensity of initiation given in expression (2) shows that estimates for ν are highly sensitive to the choice of n. Results within biologically reasonable limits will thus be obtained only for very few values n.

Growth and Mutation Rates

For $n = 1$ it has been shown in [16] that three functions of ψ are uniquely determined by $S(t|n = 1, \psi)$. Their proof can be generalized to $n \geq 1$. This is intuitively plausible. Given n, the intensity of initiation depends on the product $N_0 \nu^n$, but not on N_0 and ν individually. And the speed at which a clone of initiated cells grows depends only on the difference $\beta - \delta$, but not on the actual pair β, δ.

Lemma 2. *Let* (n, ψ) *and* $(\tilde{n}, \tilde{\psi})$ *be two sets of parameters such that* $S(t|n, \psi) \equiv_t S(t|\tilde{n}, \tilde{\psi})$. *Then we have*

$$\nu^n N_0 F_P(t;\psi) \equiv_t \tilde{\nu}^n \tilde{N}_0 F_P(t;\tilde{\psi}).$$

Proof: Let us define the integral $I(t; n, \psi) = \int_0^t \lambda_I(t-x) F_P(x)\mathrm{d}x$. This means that $S(t|n, \psi) = \exp\{-I(t; n, \psi)\}$. Note that by Proposition 1 we have $n = \tilde{n}$. So we must show that if

$$I(t; n, \psi) \equiv I(t; n, \tilde{\psi}), \tag{5}$$

then

$$\nu^n N_0 F_P(t;\psi) \equiv \tilde{\nu}^n \tilde{N}_0 F_P(t;\tilde{\psi}).$$

First, we transform $I(t; n, \psi)$ via $(n - 1)$ repeated integrations by parts into a n-fold integral. Next, we differentiate this expression n times with respect to t, to obtain

$$\frac{d^n}{dt^n} I(t; n, \psi) = n! \nu^n N_0 F_P(t;\psi).$$

Application of these two steps to both sides of (5) proves the result.

Identifiability of the Mixture Structure

Besides the parameters of the multistage model itself, we must also investigate the identifiability of the newly introduced mixture structure. Let $\mathcal{G}$ be a family of distribution functions for a certain parameter θ. Then, $\mathcal{G}$ induces the family of mixture models

$$\mathcal{S}=\left\{\int S(t\mid\theta)\mathrm{d}G(\theta);\mathrm{G}\in\mathcal{G}\right\}.$$

Family $\mathcal{S}$ is said to be identifiable with respect to $\mathcal{G}$, if

$$\int S(t\mid\theta)\mathrm{dG}_1(\theta)\equiv_t \int \mathrm{S}(t\mid\theta)\mathrm{dG}_2(\theta)\Rightarrow \mathrm{G}_1\equiv_\theta \mathrm{G}_2$$

holds for all G_1, $G_2 \in \mathcal{G}$. In other words, the population survivor function must uniquely determine the underlying mixing distribution within a pre-specified family.

This condition turns out to be hard to verify in general settings and we must focus on special cases. A very useful result was given in [20] for finite mixtures. Let Θ be a set of possible parameter values $\{\theta_1, \theta_2,...\}$ such that $\theta_1 < \theta_2 <$ .... Then the finite mixture model $S(t)=\sum_{i=1}^{g}\pi_i S(t\mid\theta_i)$ is identifiable if

$$\exists a\in\mathbb{R}\cup\infty \text{ such that } \lim_{t\to a}\frac{S(t|\theta_{i+1})}{S(t|\theta_i)}=0,\ \forall i. \tag{6}$$

This condition ensures identifiability of all finite mixtures of the survivor functions $\{S(t|\theta_i); i = 1, 2,...\}$ even without specifying the number of components g. Teicher's result requires additional regularity conditions, but these are trivially satisfied in the case of the multistage model (1).

Initiation

The multistage model we consider here describes initiation as a sequence of discrete events, namely rate limiting mutations, which lead to a cell capable of accelerated growth. A biological mechanism generating heterogeneity at this stage are germ line mutations of the genes involved, leading to individuals starting life with all cells in an intermediate stage. Mathematically, this means that the population survivor function is

$$S(t)=\sum_{i=1}^{g}\pi_i S(t\mid n_i,\psi).$$

The next proposition shows that such a mixture is identifiable.

Proposition 3. *The family of finite mixture models induced by* $\{S(t|n, \psi); n = 1, 2,...\}$ *is identifiable.*

Proof: We show that condition (6) is satisfied. The initiation incidence rates can be written recursively

$$\lambda_I(t;n+1)=\frac{n+1}{n}\nu t\lambda_I(t;n).$$

Thus, for $t > t_1 := \frac{2n}{\nu(n+1)}$,

$$\int_0^t \lambda_I(x;n)(\frac{n+1}{n}\nu x-1)F_P(t-x)\mathrm{dx}$$

$$\geq \int_0^{t_1}\lambda_I(x;n)(\frac{n+1}{n}\nu x-1)F_P(t-x)\mathrm{dx}+\underbrace{\int_{t_1}^{t}\lambda_I(x;n)F_P(t-x)\mathrm{dx}}_{:=\Lambda(t)}.$$

Since $S(t|n, \psi) \to 0$ for $t \to \infty$, we have Λ $(t) \to \infty$ for $t \to \infty$. This implies

$$\frac{S(t|n+1,\psi)}{S(t|n,\psi)}\xrightarrow{t\to\infty}0.$$

Promotion

Promotion is a complicated process and both genetic and epigenetic factors seem to be involved. Therefore, heterogeneity can be due to many different mechanisms. In the context of the multistage model, there are two main parameters these agents can influence: the growth advantage of initiated cells, γ, and the rate of malignant transformation, μ.

We can derive a result similar to the one in the previous section. Let there be a discrete set of γ-values $0 < \gamma_1 < \gamma_2 <$ ... . Note that we consider $\delta = \gamma_i - \beta_i$ as fixed, i.e. we assume in fact that there is an analogous sequence of β_i. From now on, we will write ψ for the parameter vector $(n, N_0, \nu, \delta, \mu)$.

Proposition 4. *The family of finite mixture models induced by* $\{S(t|\gamma_i, \psi); i = 1, 2,...\}$ *is identifiable.*

Proof: We will first check condition (6) in the case $\delta > 0$. We have

$$\frac{S(t|\gamma_{i+1},\psi)}{S(t|\gamma_i,\psi)}=e^{-\int_0^t\lambda_I(t-x)[F_P(x|\gamma_{i+1})-F_P(x|\gamma_i)]\mathrm{dx}},$$

and the assumption $\delta > 0$ implies that F_P is improper and converges to a limit $a(\gamma, \delta) < 1$ as t $\to \infty$. The value 1 - $a(\gamma, \delta)$ is the probability that a clone of initiated cells (generated by a single initiated cell at time $t = 0$) eventually dies out without ever giving rise to a promoted cell. The assumption $\gamma_{i+1} > \gamma_i$ implies that $a(\gamma_{i+1},\delta) > a(\gamma_i, \delta)$, and as a consequence

$$\int_0^t \lambda_I(t-x)[F_P(x\,|\,\gamma_{i+1}) - F_P(x\,|\,\gamma_i)]dx \xrightarrow{t\to\infty} \infty.$$

Let us next consider the case $\delta = 0$, and thus $\gamma_i = \beta_i$. The function F_P is in this case equal to

$$F_P(x\,|\,\beta) = \frac{\mu - \mu e^{-(\beta+\mu)x}}{\mu + \beta e^{-(\beta+\mu)x}}.$$

Using the mean value theorem we have

$$F_P(x\,|\,\beta_{i+1}) - F_P(x\,|\,\beta_i) = (\beta_{i+1} - \beta_i)\frac{\partial}{\partial\beta}F_P(x\,|\,\beta)\bigg|_{\tilde\beta},$$

where $\tilde\beta$ lies between β_i and β_{i+1}. A direct calculation shows that

1. $\frac{\partial}{\partial\beta}F_P(0\,|\,\tilde\beta) = 0\ = 0,$
2. $\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)$ is non-negative for all $x \geq 0$, and
3. $\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)$ asymptotically goes to 0 as $x \to \infty$.

Let t_0 be the (unique) maximum of $\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)$. It follows that

$$\begin{aligned}&\int_0^t \lambda_I(t-x)\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)\mathrm{dx}\\ &= \underbrace{\int_0^{t_0} \lambda_I(t-x)\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)\mathrm{dx}}_{>\lambda_I(t-t_0)\int_0^{t_0}\frac{\partial}{\partial\beta}F_P(x|\tilde\beta)\mathrm{dx}} + \underbrace{\int_{t_0}^t \lambda_I(t-x)\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)\mathrm{dx}}_{>0}.\end{aligned}$$

This shows that

$$(\beta_{i+1} - \beta_i)\int_0^t \lambda_I(t-x)\frac{\partial}{\partial\beta}F_P(x\,|\,\tilde\beta)\mathrm{dx} \xrightarrow{t\to\infty} \infty,$$

which completes the proof.

The same idea could be applied to parameter μ. Though the biological interpretation of such a frailty model would be different, technically no new issues arise, and similar results can be established.

Fitting Mixture Models

We will now apply the proposed mixture model to the human lung cancer incidence data from [10]. These authors have studied mortalities due to lung cancer in different birth cohorts of European Americans, namely those born in the 1880s, 1890s, 1900s and 1920s. The data comes in form of a vector

$$(r_i, o_i),\ i = 1,\ldots,N,$$

where r_i counts the population at risk and o_i counts the observed cancer cases during the time interval $[t_i, t_{i+1})$. The data is discussed in [21] and is publically available ([22]). Additional information is given in [23].

In our case, the data is grouped into 5-year age groups: 0–4 years, 5–9 years, 10–14 years and so on. Figures 2 and 3 show the raw hazard estimates

$$\hat\lambda_i = \frac{o_i}{r_i(t_{i+1} - t_i)}.$$

As mentioned earlier, the observed hazard has a peak at around 80 years and a decrease for the higher ages, while the hazard of the multistage model given by (1) is strictly monotonic increasing. Estimation of the parameters by analytic graduation as described later on leads to extremely poor fits, which for all ages above 30 and for all birth cohorts give quite useless predictions. The fault does, however, not lie with the methods of estimation but rather with the model. Thus, using the inverse of the vari-

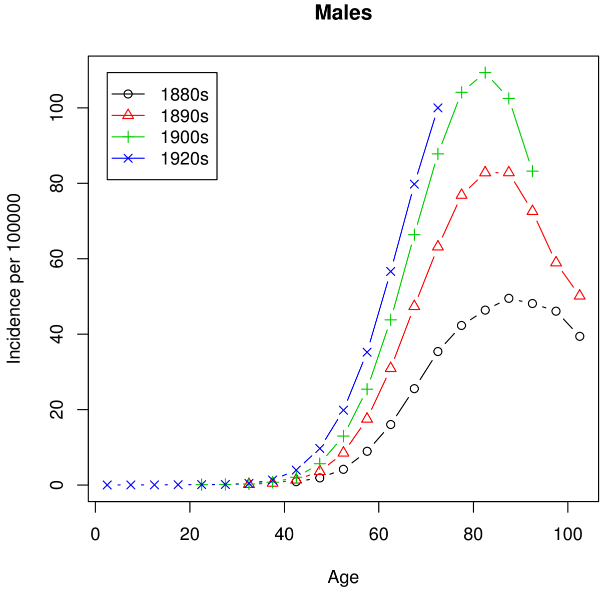

Figure 2
Observed incidence (males). Observed lung cancer incidence rates in the United States for four birth cohorts. The population considered are the males of European descent.

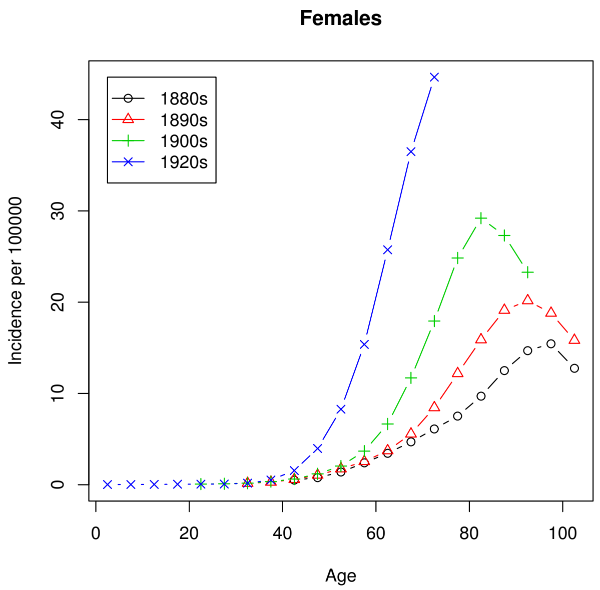

Figure 3
Observed incidence (females). Observed lung cancer incidence rates in the United States for four birth cohorts. The population considered are the females of European descent.

ance of the estimated incidence rates as weights leads to almost the same poor fit. The unmixed multistage model does not succeed in describing the incidence rates in Figures 2 and 3. We will come back to this failure.

Two component mixtures on the other hand are flexible enough to provide good fits. We will illustrate this using the γ-frailty model

$$S(t) = \pi_l S(t|\gamma_l) + \pi_u S(t|\gamma_u), \qquad (7)$$

where $0 \le \pi_l \le 1$, $\pi_l + \pi_u = 1$, $0 < \gamma_l < \gamma_u$. To get identifiability, we will fix N_0 and δ. But in order to get stable estimates, we fix also γ_l and n. Note that the parameters we estimate have a restricted domain of definition,

$$(\pi_l, \nu, \gamma_u, \mu) \in (0,1) \times \mathbb{R}_+ \times (\gamma_l, \infty) \times \mathbb{R}_+.$$

We will use suitable transformations to respect these constraints.

Maximum Likelihood Estimation

We treat the failures from competing causes as right censorings. This means for each time interval $[t_i, t_{i+1})$ we observe o_i failures due to cancer and have $c_i = r_i - r_{i+1} - o_i$ censored individuals. Under the assumption of independent and uninformative censoring, the likelihood function $L(\pi_l, \nu, \gamma_u, \mu | N_0, \delta, n, \gamma_l)$ is given by

$$\prod_{i=0}^{N} \left(S(t_i) - S(t_{i+1}) \right)^{o_i} S(t_i)^{c_i}.$$

By numerically optimizing this likelihood, we observe a strange behavior of the MLE. Figure 4 shows the data from the males 1880s cohort along with the models corresponding to the MLE, the least squares fit LSE, and the starting value of the numerical optimization. As we can see, the MLE fails completely to catch the behavior of the observed incidence at old ages; only the first few data points are well fitted. Convergence to this model seems even more astonishing when we consider the initial model. The chosen starting value is far away from the data in terms of fit, but it is close to the observed hazard in terms of shape. Furthermore, the model corresponding to the LSE fits the observed hazard very closely. This shows that the parametric family we apply to the data does indeed contain models that can fit. But in this example, likelihood and fit do not measure the same thing. The huge discrepancy, however, is intriguing. The strange behavior of the MLE is caused by several effects. One aspect is model mis-specification in relation with the special metric used in likelihood based inference. The data is not really generated by our multistage model, while the MLE corresponds to the survivor function that minimizes

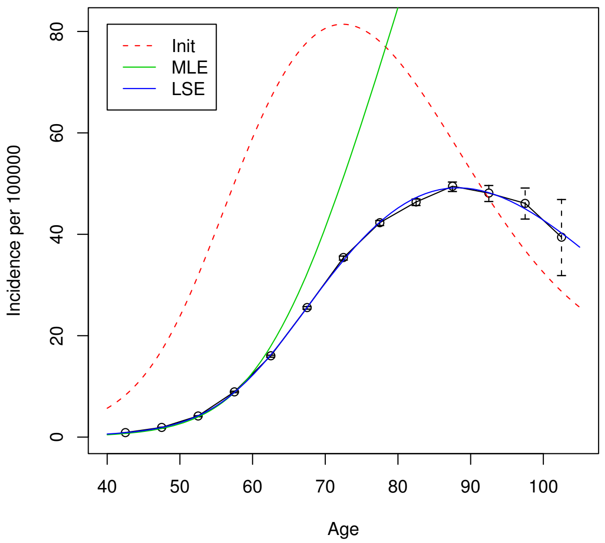

Figure 4
ML and LS fits. Lung cancer incidence rates for the European American males born in the 1880s. The superposed curves show the fitted hazards of the carcinogenesis model (7) based on the MLE and least squares. In the fitting process, $N_0 = 10^{10}$, $\delta = 0$, $n = 2$ $\gamma_l = 10^{-4}$ were kept constant. The initial value for the remaining parameters were $\pi_l = 0.97$, $\gamma_u - \gamma_l = 0.2$, $\nu = 10^{-6.5}$ and $\mu = 10^{-5}$.

the Kullback-Leibler distance to the observed empirical survivor function. But this is a very special metric and can produce obviously strange results in some cases.

In mechanistic modeling, likelihood based inference is often difficult due to local maxima and/or low curvature around the maxima. Both problems apply to our case. Our likelihood-surface is multimodal because the different biological parameters compete. This problem can be avoided by extensive use of the available biological knowledge. If we have good starting values and restrict attention to biologically reasonable intervals, then the likelihood surface is unimodal in that domain. The second problem is more difficult to treat. Even for identifiable parameters the likelihood surface is often extremely flat around its maximum. Figure 5 gives the contour plot of the log-likelihood for a reduced parameter space. That is, we take model (7), but fix all parameters except ψ = (logitπ_l, $\log_{10}\mu$). The log-likelihood essentially has a ridge starting the upper-right corner and running downwards as one moves to the left. This means that only a combination of the two parameters can be estimated precisely, but not both separately. The log-likelihood values of the estimates in Figure 5 are $l(\hat{\psi}_{\mathrm{ML}})$ = -1.338 · 10^6 and $l(\hat{\psi}_{\mathrm{LS}})$ = -1.355 · 10^6. While these values appear to be close, they are in fact quite different in the likelihood metric, because

$$2(l(\psi_{\mathrm{ML}}) - l(\psi_{\mathrm{LS}})) \simeq 32600 \gg q\chi_2^2(0.999) \approx 13.8.$$

A 95% confidence region determined by a likelihood-ratio test is shown in Figure 6. Note how small this confidence set is. So, not only does the likelihood technology give badly fitting hazard rates, it is also overly optimistic about having found the right values.

The most important reason that leads to the failure of the MLE in our application, however, is the heavy censoring. We deal with human cancer incidence data. This means we consider a rare event, and most members of the population fail from competing causes. In the data set we are considering there are tens of millions individuals at risk at the first time points, but only some tens of thousands at the last one. In order to illustrate the impact of censoring, we will construct a sequence of artificial data sets that lead to the same raw hazard estimates, but differ in the degree of censoring. As before, we note by (r_i, o_i) the real data set. Let us define the points ($\tilde{r}_i^k$, $\tilde{o}_i^k$) by

$$\tilde{r}_i^k = 10^6 - i \cdot k 10^4, \quad \text{and} \quad \frac{\tilde{o}_i^k}{\tilde{r}_i^k} = \frac{o_i}{r_i}. \qquad (8)$$

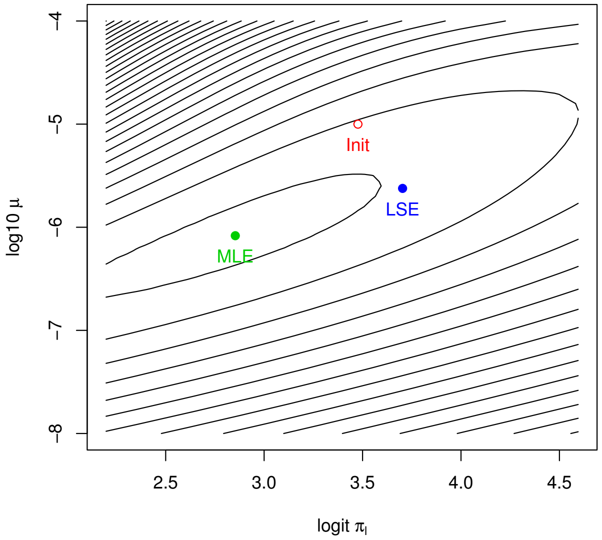

Figure 5
Log-likelihood surface. Contour plot of the log-likelihood surface. The parameter space is reduced by keeping all the parameters except two constant. The two parameters shown are logitπ_l and $\log_{10}\mu$.

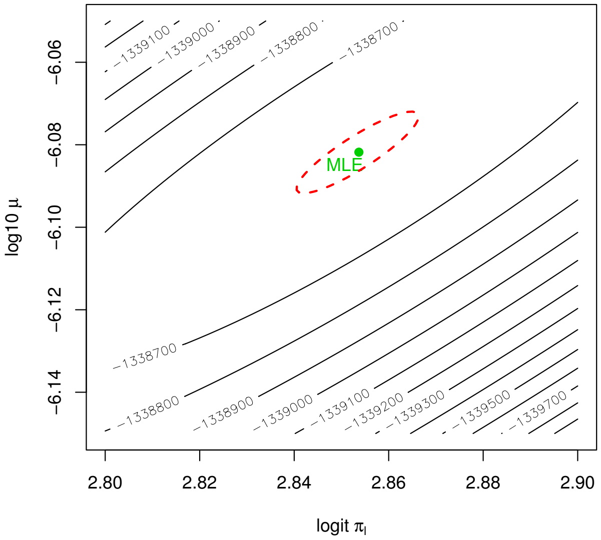

Figure 6
Log-likelihood surface (zoomed). Contour plot of the log-likelihood surface as shown in Figure 5. The plot zooms in on the MLE and in addition contains a 95% confidence ellipsoid.

That is we start with a population of size 10^6 and suppose that during every time interval exactly $k10^4$ individuals die – either due to cancer or due to competing causes. We then fit model (7) by maximum likelihood as before (we consider again the four parameters π_l, ν, γ_u, μ as unknown). Figure 7 gives the estimated models for k = 1,...,8. The MLE behaves better for small k than for large k. In Figure 8 we calculated the residual sums of squares (RSS) for these models, which seem to increase exponentially with the coefficient of variation of the $\tilde{r}_i^k$ sequence,

$$\mathrm{cv}_k = \frac{\mathrm{sd}(\tilde{r}_0^k,\ldots,\tilde{r}_{12}^k)}{\mathrm{mean}(\tilde{r}_0^k,\ldots,\tilde{r}_{12}^k)}.$$

The above example shows that the MLE is dominated by the points corresponding to large "at risk" sets. The LSE, on the other hand, works fine, since it attributes equal weight to all age intervals. This makes one wonder whether a weighted least squares approach would suffer from the same problem as the MLE. If we give for example weights proportional to the population at risk, would the LSE break down as well? The answer to this question is clearly no. Considering Figure 4 once again, we realize that there is a model that fits all the data points very accurately. This model will be good even if we downweight the contribution to the RSS of the points at high ages. Any weighted least squares approach will select a model that is very close to the standard LSE.

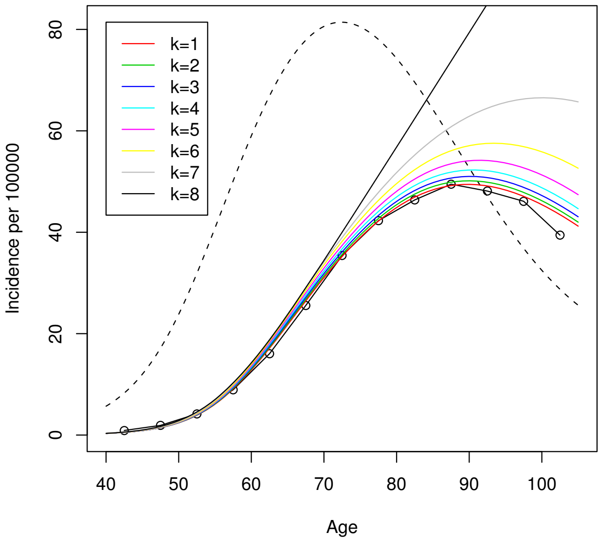

Figure 7
ML fits to artificial data. The hazard curves corresponding to the maximum likelihood fits for the data sets constructed according to (8) for different values of k.

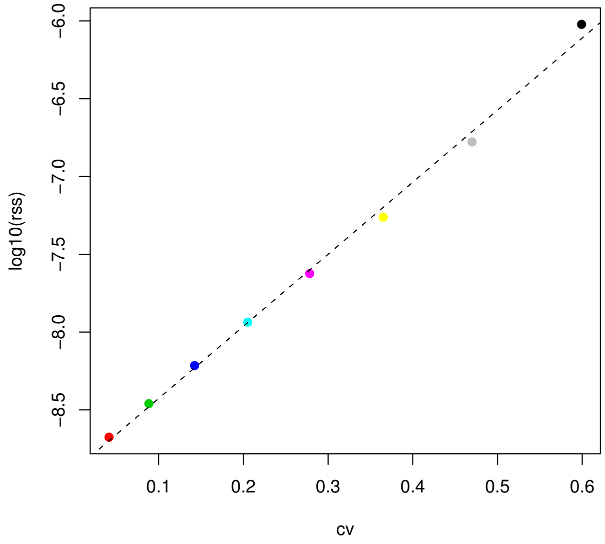

Figure 8
RSS of ML fits. The logarithm of the residual sums of squares of the various fits shown in Figure 7 as a function of the coefficient of variation of the size of the at risk set. Note that for the real data set we have cv(r_0,...,r_{12}) ≈ 0.77.

Analytic Graduation

The LS estimates shown in the previous figures were obtained by analytic graduation, which is a standard procedure to fit continuous curves to discretized data. A detailed discussion of the procedure and derivation of asymptotic results can be found in [24,25].

Figures 9 and 10 show model (7) fitted to the different cohorts. The model successfully reproduces the observed data. Table 1 gives the corresponding parameter estimates. Note that these values are conditional given N_0, δ, γ_l and n. The value N_0 acts as a scale parameter. Changes in N_0 are compensated by ν such that the product $N_0\ \nu^n$ stays more or less constant. The other parameters also remain quite stable. The effect of δ is rather fuzzy, no clear conclusions emerge. In all cases where enough data at old ages is available (i.e. all but the 1920s cohort), the estimated proportion of the population at high risk, $\hat{\pi}_u$, is not sensitive to changes in the fixed parameter values. The peak of the observed hazard determines $\hat{\pi}_u$ quite accurately. Finally, reasonable results can also be obtained for n = 3, while other choices of n produce unrealistic estimates for at least some of the parameters. Note that good fits can be

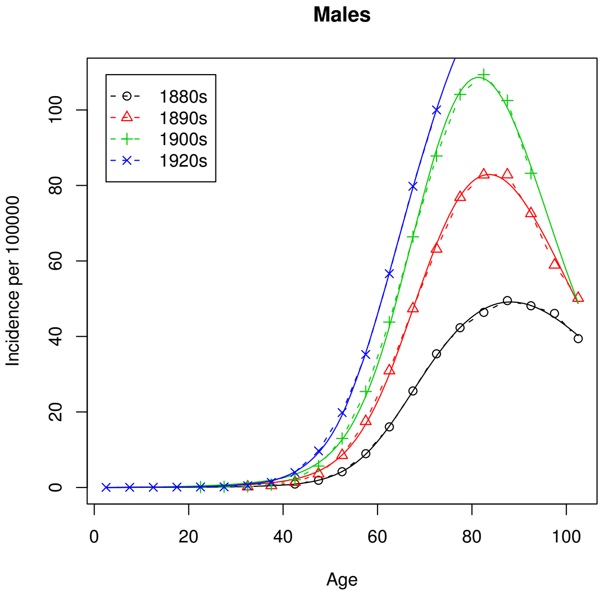

Figure 9
LS fits (males). Observed (dashed lines) and modeled (solid lines) incidence rates for the data from Figure 2.

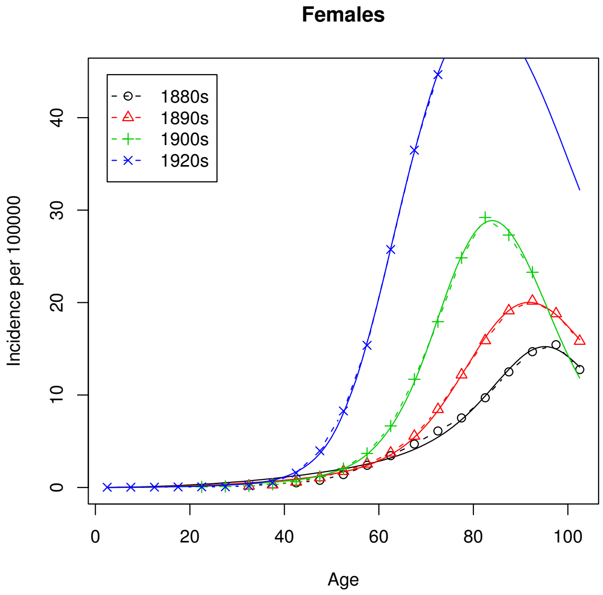

Figure 10
LS fits (females). Observed (dashed lines) and modeled (solid lines) incidence rates for the data from Figure 3.

achieved only as long as γ_l is small enough. We set $\gamma_l = 10^{-4}$ in the models given here.

The least squares estimates for a single unmixed multistage model in which all the parameters except N and δ are fitted, lead to curves ressembling the maximum likelihood fit in Figure 3. The estimated parameters show that the simple multistage model attempts to distinguish between the earlier and later cohorts to a large extent by increasing the initiation rate ν. The increase is three-fold between 1880 and 1920 for the females and even six-fold for the males. The other parameters remain more or less constant across the cohorts. The incidence rates in females is much lower than in males. While the mixture model adjusts to this through an adjustment of the mixture weights, the single multistage model explains it by very different estimates of the promotion rate μ between the sexes.

In order to assess the accuracy of the given estimates, we use projections of a joint confidence region rather than marginal confidence intervals. In other words, we determine a confidence region $C \subset \mathbb{R}^4$, and then we look at projections of C on the six parameter plains spanned by the four parameter axis. The confidence region we get for the EAMs 1880s cohort is shown in Figure 11. The confidence region reveals the strong dependencies between the different parameters. Though a parametrization with such dependant parameters is unsatisfactory from a mathematical point of view, the dependencies might be interesting in biological terms. Not the two mutation rates ν and μ seem to compete, but rather the net growth rate γ and the two mutation rates. So at the extremes of the confidence region, we have models with high mutation rates but low proliferation of initiated cells, or models with low mutation rates but large cell growth. Note that the corresponding hazard curves are markedly different.

Table 1: Conditional parameter estimates given the fixed values n = 2, $N_0 = 10^{10}$, $\delta = 0$ and $\gamma_l = 10^{-4}$.

Cohort	Males				Females			
	$\hat{\pi}_u$	$\hat{\nu}$	$\hat{\gamma}_u$	$\hat{\mu}$	$\hat{\pi}_u$	$\hat{\nu}$	$\hat{\gamma}_u$	$\hat{\mu}$
1880s	0.021	2.5×10^{-7}	0.183	3.5×10^{-6}	0.003	4.9×10^{-7}	0.134	2.2×10^{-6}
1890s	0.029	3.2×10^{-7}	0.173	4.4×10^{-6}	0.005	4.3×10^{-7}	0.146	2.1×10^{-6}
1900s	0.034	3.6×10^{-7}	0.167	5.2×10^{-6}	0.007	4.8×10^{-7}	0.168	1.3×10^{-6}
1920s	0.077	1.9×10^{-7}	0.189	7.5×10^{-6}	0.023	2.4×10^{-7}	0.203	3.2×10^{-6}

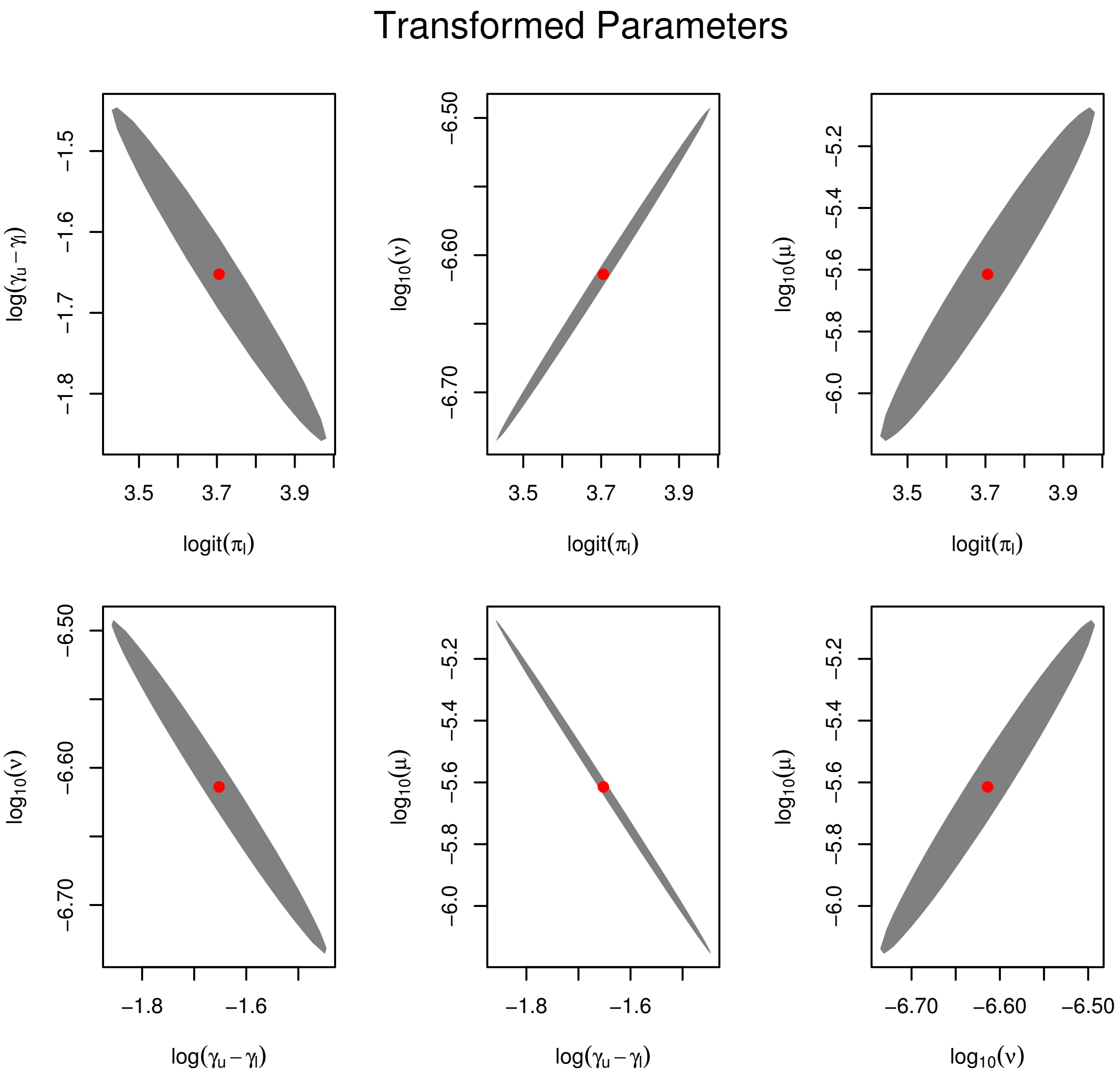

Figure 11

95% asymptotic confidence region. Projections of an asymptotic 95% confidence region (CR) for the parameters of the carcinogenesis model fit the the 1880s birth cohort of European American males. Note that we used the parametrization (logit(π_l), log(γ_u - γ_l), log$_{10}$ ν, log$_{10}$ μ) in the fitting process.

Discussion

The γ-frailty model we fitted to our data is such that only one parameter involved in promotion is allowed to vary between population subgroups. Such a model would suggest a process where initiated cells are created in all individuals according to the same dynamics, but only in a small subgroup of the population promotion and malignant transformation happen. This is consistent with the fact that promotion depends on stimuli that might be present only in a fraction of the population. The proportion of high risk individuals, estimated by $\hat{\pi}_u$, reflects the change of the hazard curves between the 1880s and the 1920s cohorts. The sharp increase of maximal incidence in a relatively short period of time must be due to environmental factors such as occupational exposure and smoking.

We got satisfactory fits only when we allowed for two clearly separated population subgroups with a low risk group that runs a risk close to zero. This is consistent with the results reported by the other research groups that introduced frailty into carcinogenesis modeling. In [10] the estimated fraction at risk is very low. And also in [8] the estimated proportion of susceptibles is lower than

0.5%, though these authors worked with Scandinavian data on testicular cancer.

Besides the γ-frailty model we considered here, one could clearly build mixture models using other parameters. In particular the number of mutations for initiation, n, the rate of malignant transformation, μ, or the initiating mutation rate ν would be natural choices. The increase in lung cancer rates that was observed during the 20thcentury coincides with an increase in the fraction of smokers. Smoking might very well influence μ and ν. However, when applying such models to our data, no new issues arise, and we omit a detailed discussion here. Still, one needs to realize that all the mentioned models can fit the data equally well. This is not surprising, since we fit a relatively complex model to a very simple data structure. However, in all approaches we tested, the data suggested two component mixtures with a small high risk group and a large quasi immune group.

Conclusion

We have studied an extension of the multistage carcinogenesis model by mixture. This allowed us to introduce population heterogeneity. The multistage model is a mechanistic model and all its parameters have a biological interpretation. Therefore, it is natural to introduce the notion of frailty in a biologically meaningful way. Such an approach is given by our mixture models, which can reproduce observed human lung cancer incidence data very accurately. Very good fits are achieved with very simple, two component mixtures also in cases where a continuous distribution might seem more adequate. However, the peak observed in the population hazard rates can be reproduced by continuous mixture models only when the population is clearly separated into a high risk and a low risk group. In other terms, the density of such a distribution would typically be bathtub shaped, closely resembling two component mixtures. Biological systems are often buffered. Small changes in the environment have no significant effect, and only after passing over some threshold value may abrupt changes in the system occur. Since we consider a late end-point, namely cancer, in a very complicated system, it is not surprising that we obtain in Section 4 mixture models that reflect such a buffering. It would be interesting to link the model with concrete biological mechanisms that are able to explain flip-flop processes. This would be an approach to understand how heterogeneity acts upon human carcinogenesis.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank the editor and the referees for thoughtful comments and their help in improving the paper. The first author has been supported by a grant from the Swiss National Science Foundation.

References

1. Nordling CO: **A new theory on the cancer-inducing mechanism.** *Brit J Cancer* 1953, **7:**68-72.
2. Armitage P, Doll R: **The age distribution of cancer and a multistage theory of carcinogenesis.** *Brit J Cancer* 1954, **8:**1-12.
3. Moolgavkar SH, Venzon DJ: **Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors.** *Mathematical Biosciences* 1979, **47:**55-77.
4. Moolgavkar SH, Knudson A: **Mutation and cancer: A model for human carcinogenesis.** *J Nat Cancer Inst* 1981, **66:**1037-1052.
5. Kopp-Schneider A: **Carcinogenesis models for risk assessment.** *Statistical Methods in Medical Research* 1997, **6:**317-340.
6. Edler L, Kitsos CP, (Eds): *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment* Wiley Series in Probability and Statistics, West Sussex, England: John Wiley & Sons; 2005.
7. Tan WY, Singh KP: **A mixed model of carcinogenesis with applications to retinoblastoma.** *Mathematical Biosciences* 1990, **98:**211-225.
8. Aalen OO, Tretli S: **Analyzing incidence of testis cancer by means of a frailty model.** *Cancer Causes and Control* 1999, **10:**285-292.
9. Moger TA, Aalen OO, Halvorsen TO, Storm HH, Tretli S: **Frailty modelling of testicular cancer incidence using Scandinavian data.** *Biostatistics* 2004, **5:**1-14.
10. Morgenthaler S, Herrero P, Thilly WG: **Multistage carcinogenesis and the fraction at risk.** *Journal of Mathematical Biology* 2004, **49(5):**455-467.
11. Kopp-Schneider A, Portier CJ, Sherman CD: **The exact formula for tumor incidence in the two-stage model.** *Risk Analysis* 1994, **14(6):**1079-1080.
12. Zheng Q: **On the exact hazard and survival functions of the MVK stochastic carcinogenesis model.** *Risk Analysis* 1994, **14(6):**1081-1084.
13. Moolgavkar SH, Luebeck G: **Two-event model for carcinogenesis: Biological, mathematical, and statistical considerations.** *Risk Analysis* 1990, **10(2):**323-341.
14. Hoogenveen R, Harvey J, Andersen M, Slob W: **An alternative exact solution of the two-stage clonal growth model of cancer.** *Risk Analysis* 1999, **19:**9-14.
15. Heidenreich WF: **On the parameters of the clonal expansion model.** *Radiat Environ Biophys* 1996, **35:**127-129.
16. Hanin LG, Yakovlev AY: **A nonidentifiability aspect of the two-stage model of carcinogenesis.** *Risk Analysis* 1996, **16(5):**711-715.
17. Heidenreich WF, Luebeck EG, Moolgavkar SH: **Some properties of the hazard function of the two-mutation clonal expansion model.** *Risk Analysis* 1997, **17(3):**391-399.
18. Sherman CD, Portier CJ: **The two-stage model of carcinogenesis: overcoming the nonidentifiability dilemma.** *Risk Analysis* 1997, **17(3):**367-374.
19. Luebeck EG, Moolgavkar SH: **Multistage carcinogenesis and the incidence of colorectal cancer.** *PNAS* 2002, **99(23):**15095-15100.
20. Teicher H: **Identifiability of finite mixtures.** *Annals of Mathematical Statistics* 1963, **34:**1265-1269.
21. Herrero-Jimenez P: **Determination of the historical changes in primary and secondary risk factors for cancer using U.S. public health records.** In *PhD thesis* MIT; 2001.
22. **MIT epidemiology database pages** [Http://epidemiology.mit.edu/]
23. Herrero-Jimenez P, Tomita-Mitchell A, Furth EE, Morgenthaler S, Thilly WG: **Population risk and physiological rate parameters for colon cancer: The union of an explicit model for carcinogenesis with the public health records of the United States.** *Mutation Research – Fundam Molec Mechan Mutagenesis* 2000, **447:**73-116.
24. Hoem JM: **On the statistical theory of analytic graduation.** In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Theory of statistics Volume I.* Berkeley, Calif.: Univ. California Press; 1972:569-600.
25. Hoem JM: **The statistical theory of demographic rates. A review of current developments.** *Scand J Statist* 1976, **3(4):**169-185. [With discussion by Niels Keiding, Hannu Kulokari, Bent Natvig, Ole Barndorff-Nielsen, Jørgen Hilden and a reply by the author].