HETEROGENEITY IN MULTISTAGE CARCINOGENESIS AND MIXTURE MODELING

THÈSE N^{O} 3611 (2006)

PRÉSENTÉE LE 8 SEPTEMBRE 2006

À LA FACULTÉ SCIENCES DE BASE

Institut de mathématiques

SECTION DE MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Sandro GSTEIGER

mathématicien diplômé de l'Université de Fribourg
de nationalité suisse et originaire de Grindelwald (BE)

acceptée sur proposition du jury:

Prof. T. Mountford, président du jury
Prof. S. Morgenthaler, directeur de thèse
Prof. A. C. Davison, rapporteur
Dr A. Kopp-Schneider, rapporteur
Prof. L. Preziosi, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Lausanne, EPFL
2006

Acknowledgements

It is a long-term project to write a PhD thesis, and I am indebted to many people who have contributed, consciously or not, to its successful achievement.

First of all, I am deeply grateful to Prof. Stephan Morgenthaler, my supervisor, for his support and encouragement, without which this thesis would not have been accomplished. I thank PD Dr. Annette Kopp-Schneider and Prof. Luigi Preziosi for the interest they have shown in my work as external experts and, in the case of the latter, also as a host. Furthermore, thanks go to the internal expert, Prof. A. Davison, and to the director of the jury, Prof. T. Mountford.

During the whole period at EPFL my colleagues gave important assistance, and I express my thanks to all present and former graduate students of the two statistics chairs. In particular, I would like to thank Andrei, Thomas, Baptiste, Jérôme, Marc-Olivier, and Lionel for all their help, but also for their humour, their brightness, and their friendship.

Finally, I thank my family and friends, who gave me constant and unconditional support. In particular, I am indebted to Martin, David, Bernhard, Stefan, Giovanni, and Mirko, and of course my thanks go to Katrin, who contributed with all her enthusiasm, her serenity, and her humanity throughout the years I was working on this thesis. The help of all these people cannot be valued enough.

Abstract

Carcinogenesis is commonly described as a multistage process. In a first step, a stem cell is transformed via a series of mutations into an intermediate cell having a growth advantage. Under favorable conditions, such a cell will give rise to a clone of initiated cells. Eventually, further alterations may transform a cell out of this clone into a malignant tumor cell.

A mechanistic model of this process is given by the widely used two-stage clonal expansion model (TSCE). In this thesis, we take up a generalization of the TSCE, and study, how to introduce the concept of population heterogeneity into the model. We use mixture modeling, which allows to describe frailty in a biologically meaningful way.

In a first part, we focus on theoretical properties of the extended model. Especially identifiability is discussed extensively. In a second part, we fit the model to human cancer incidence data. We analyze a situation, in which maximum likelihood estimation fails, and describe alternatives for statistical inference. The applications show that good fits are achieved only when the mixing distribution separates the population clearly into a large, virtually immune group, and into a small, high risk group.

Keywords: Multistage carcinogenesis; Heterogeneity; Frailty modeling; Mixture modeling

Version Abrégée

De manière générale, la carcinogénèse est décrite comme un processus à étapes multiples. Dans une première phase, une cellule souche subit des mutations qui ont l'effet de lui donner la capacité d'une croissance accélérée. Sous des conditions favorables, une telle cellule donnera lieu à un clone de cellules initiées. Dans une deuxième phase, une cellule initiée est transformée en cellule de cancer.

Le modèle à deux étapes avec expansion clonale (DEEC) est un modèle méchaniste souvent utilisé pour modéliser la carcinogénèse. Le but de cette thèse consiste à introduire la notion d'hétérogénéité dans le modèle DEEC. On va utiliser les modèles mélangés à cette fin. Cette méthodologie nous donne la possibilité d'introduire la notion de fragilité d'une manière à avoir un sens biologique.

Nous discutons d'abord les propriétés théoriques du modèle. En particulier, nous étudions en détail l'identifiabilité. Ensuite, nous montrons des applications du modèle à des données d'incidence chez l'homme. Nous sommes confrontés à une situation, dans laquelle l'estimation par maximum de vraisemblance échoue, ce qui nous amène à proposer des alternatives. Les applications suggèrent une population qui consiste de deux groupes, dont un est à risque très bas.

Mots clés: Carcinogénèse; Modèle à étapes multiples; Hétérogénéité; Modèles de fragilité; Modèles mélangés

Contents

CHAPTER 1

Introduction

Sadly, we are all familiar with cancer. In the more developed countries, it is the second leading cause of death behind cardiovascular disease. But it afflicts all communities worldwide. We are all confronted with cancer either directly, as a patient, a friend or a relative of a patient, or indirectly, through campaigns and debates. And although prevention and treatment do show successes, the burden of cancer will tend to increase in countries, in which the age structure of the population shifts towards a larger proportion of older people.

This thesis is situated in the field of mathematical cancer research. We work with a class of carcinogenesis models that are based on biological theory. The reader need not be frightened, however, because we start with a short introduction to cancer biology, and will also review some statistical methodology that will be used extensively later on.

1.1 Basic Cancer Biology

The term cancer does not stand for a single disease, but represents a collection of diseases characterized by uncontrolled cell proliferation. Cells that would be quiescent in their normal state continue cell division, form clones, eventually invade surrounding tissues and metastasize to other parts of the body through blood vessels or lymphatic channels. A short list of basic definitions that will be used extensively in this text is as follows:

Tumor General term for an uncontrolled growth of cells.

Neoplasm Same as tumor.

Benign tumor A tumor that does not metastasize or invade surrounding tissues.

Malignant tumor A tumor that has the ability to metastasize or invade surrounding tissues.

Cancer Same as malignant tumor.

Metastasis Ability to establish secondary tumor growth at a new location in the body away from original site.

Carcinogenesis Formation of a carcinoma. *Or:* General term for the formation of any type of cancer.

Cancer can arise in virtually any tissue. But it is much more likely to develop in frequently renewing cells, like epithelial cells, than in parts of the body that do not normally proliferate, like neurons. Malignant tumors can be classified according to their tissue type of origin. Carcinomas, the most common form of cancer, arise in epithelium (sheets of tightly packed cells that line organs and body cavities). Sarcomas arise in connective or muscle tissues. Leukemia and lymphoma are malignant tumors of the hematopoietic system, the blood forming structure of the body, and have no benign counterparts.

The term carcinogenesis is used in two different ways. The first one refers to the production of a carcinoma (epithelial cancer). The second one means the general process leading to any type of tumor. Throughout this text, we will refer to the latter meaning, when talking about carcinogenesis.

Although clinically cancer is a large group of diseases that vary in age of onset, growth rate, invasiveness, potential to form metastasis, treatability, etc., the molecular and cellular mechanisms that lead to cancer remain similar for most cancer types. For general accounts on cancer biology see for example Ruddon (1995); Franks and Teich (2001). See Weinberg (1996) for a general review of carcinogenesis, our next topic.

The normal, healthy body of an adult consists of more than 10^{13} cells. Some organs, like the skin, are renewed constantly while others renew very slowly. Regulation of cell proliferation is therefore a complex and crucial mechanism for the stable functioning of the body. Normal cells divide only when instructed to do so. Tumor cells violate this rule. They either proliferate without stimulation by external signals, or they become deaf to inhibitory factors.

Cancer is a genetic disease. The failure of cell cycle control is caused by mutations in certain groups of genes and several events are needed to transform a normal cell into a tumor. Proto-oncogenes code for proteins that stimulate cell division (mitosis). In mutated forms, called oncogenes, they overexpress these proteins leading to excessive cell proliferation. Their counterparts are tumor suppressors: genes which inhibit mitosis. Neighbouring cells of a non-normal dividing cell will send "stop" signals to activate those genes. Mutations in tumor suppressor genes can cause break-down of this control. A further backup system consists in programmed cell death (apoptosis). If essential parts of a cell get damaged, a signal to commit suicide is sent. Cancer cells escape also from this program. A final control mechanism is cell ageing (see Mathon and Lloyd (2001)). In normal cells specific DNA segments, called telomeres, at chromosome ends shorten a bit at every cell division. This molecular counting device is used to instruct the cell after

an intrinsically defined number of doublings to stop growth and enter into a senescent state. Cancer cells on the contrary systematically replace the telomeric segments cut off at each cell cycle. Their telomeres remain intact and cells reach immortality. A related topic is stem cell biology, which is important for therapy development. See for example Pardal et al. (2003); Jones et al. (2004).

Hallmarks of malignancy are invasiveness and ability to metastasize. A primary tumor (of immortal cells) can often relatively easily be removed surgically. But the spread of malignant cells all over the body is what makes cancer so lethal. Normal cells have structures on their surface that identify where in the body the cell should be. Such area codes mediate cell-cell adhesion and anchorage of tissues to adjacent structures. In cancer, genetic events have caused this system to go wrong (a review of cancer spread is given by Ruoslahti (1996)).

The origin of a cancer is monoclonal. A single cell can give rise to a malignant tumor through an evolutionary process. A sequence of mutations transforms a stem cell into a so-called initiated cell having the potential of a growth advantage. Under favorable conditions, the initiated cell will give rise to a clonal expansion. Proliferation of such initiated cells is facilitated by site specific promoters and may even depend completely on the presence of such agents. Examples are hormones, anabolic steroids, but also salt or tobacco. Further changes transform a cell out of this clone into a malignant tumor cell. Genetic and epigenetic events are involved in this step. The process of carcinogenesis is drawn schematically in Figure 1.

Such an initiation-promotion pattern holds for most cancer types, though the number and type of mutations involved is site specific (see Knudson (2001)). When talking about the number of mutations needed to cause cancer, what we are really interested in is the number of rate limiting events. Malignant cells show thousands of mutations that are an effect but not a cause of cancer. The role of genomic instability in the transformation of a normal cell into a cancer cell is part of ongoing debate. Many cancers show chromosomal instability, that means an increased rate of loss or gain of whole chromosome parts during cell di-

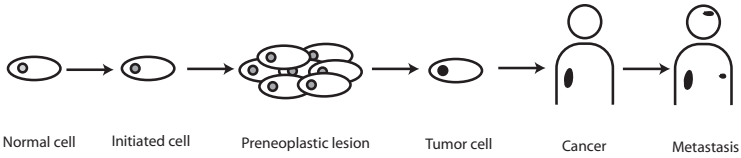

Figure 1: *Evolution of a normal stem cell into a cancer cell. Mutations lead to an initiated cell, which gives rise to a pre-neoplastic clone. A cell out of this clone may mutate further and generate a clinically detectable cancer. Finally, cancer cells spread through the body and form distant malignant clones.*

vision. This might for example accelerate the rate of tumor suppressor gene inactivation and thereby reduce the number of rate limiting steps required (see Tomlinson et al. (2002); Michor et al. (2004)).

A consequence of this theory is that carcinogenesis is intrinsically a random process. But the probability to develop cancer varies considerably among individuals. This is due to inherited factors as well as to different exposure to environmental and industrial carcinogens. Since cancer is caused by mutations, carcinogenic is basically anything that is mutagenic and anything that stimulates the rate of mitosis. Lifestyle is a major cause, the most prominent lifestyle factors being diet and tobacco use. Finally, the development of cancer can be initiated by other elements like certain viruses or chronic inflammation.

Mutations are rare events - at least in a healthy body under normal environmental conditions. Since carcinogenesis requires multiple mutations, the waiting time for tumor onset is generally several decades. Therefore, cancer is mainly a disease of older age groups. Still childhood cancer does occur. Many of such cases are explained by inherited mutations in cancer causing genes. If mutations occur in the germ line of a parent, the defective gene may be passed to the next generation. The offspring starts life with all cells in some intermediate stage. In the worst case a single mutation could be enough to cause cancer.

1.2 Mathematical Cancer Research

Quantitative applications can be found in many fields related to cancer research, and the mathematical tools used vary considerably. We give here a short overview of some selected topics with no claim of completeness.

An important field is cancer epidemiology. Based on case-control studies association between exposure to risk factors and disease can be investigated. This technique applies directly to human beings, which is a major advantage. The impact of environmental factors or lifestyle (as for example smoking) can be assessed based on observations of human populations. No species barrier has to be overcome. But case-control studies are very susceptible to bias and must be carried out carefully. One of the first such work was Doll and Hill (1950), who associated cigarette smoking to lung cancer. An extensive discussion of the use of case-control studies in cancer research give Breslow and Day (1980). Note that long term surveys become only now available; Doll et al. (2004) for example give the first long term study (50 years) of mortality in relation to smoking. A related field is disease mapping, where cancer incidence is examined for population subgroups and different geographic regions.

Once a risk factor is identified, the dose-response relationship of the corresponding agent is of main interest. Regulatory agencies wish to define threshold values such that exposure to lower concentrations of the given carcinogen is considered safe. In this process, very common issues are the extrapolation from high doses to low doses, from laboratory animals to humans, and between different routes and patterns of exposure. A very recent account of the state of the art in cancer and human health risk assessment is given in Edler and Kitsos (2005).

Fundamental to the above topic is the understanding of the cancer causing mechanisms. Quantitative descriptions of the process that transforms normal cells into cancer cells are provided by (stochastic) carcinogenesis models. Generally, those models are linked stochastic processes counting numbers of cells of different types present at a given

time. The level of biological detail to be modelled determines the type of theory to be used, for example Markov chains and branching processes. If a carcinogenesis model is mechanistic, i.e. if it is the formalization of the biological theory of cancer, then the model can be used to test biological hypothesis about the process. Consequences of quantitative models can thereby lead to insights into the biological mechanisms for tumor onset. An excellent review of stochastic carcinogenesis models is given by Kopp-Schneider (1997).

Another important topic is the modeling of cancer growth and evolution. Probabilistic models based on interacting-particle systems (contact processes) and random sets have been proposed (see Cressie (1991)). The aim is usually to derive asymptotic properties like convergence to some shape or conditions leading to extinction. Another very active research field is deterministic modeling of cancer growth, where differential methods and numerical simulation are used. Very often, this type of research is performed in close collaboration with clinicians. The better understanding of cancer growth should help to improve current treatments and can propose new strategies for therapy. See Preziosi (2003) for the state of current research in this field.

Finally, we should mention the very active fields of microarray data analysis, data mining and machine learning. The recent revolutions in molecular biology and biotechnology have had a great impact on cancer research. Many laboratories investigate patterns in gene expression of cancer tissues. Expression patterns are clustered into groups, and new cancer causing genes are identified. The enormous amount of data collected gives rise to exciting opportunities in statistical research.

1.3 Short Review of Statistical Tools

In this section, we briefly introduce some selected statistical concepts. The aim is to define the main methods that we will use later on. The topics chosen are general issues in statistics and are not restricted to cancer modeling. The reader with a solid statistical background will already be familiar with them.

1.3.1 Survival Analysis

Stochastic carcinogenesis models describe the waiting-time for tumor onset. The main object of study is the random time T we wait (in our case in years from birth) until some event of interest (in our case cancer) occurs. The statistical field that studies such problems is called survival analysis (other frequently used names are failure time data analysis, lifetime data analysis, or time to event data analysis). Examples of events considered might be infection with a disease in epidemiology, breakdown of a machine in industry, or the beginning of a new work contract for a person on unemployment as studied in the social sciences. Clearly, the notion of time is central in survival analysis and deserves special attention. Time origin and scale of a study must be selected carefully. Continuous or discrete models for time can be appropriate, and one is not restricted to physical or chronological time. For example in biology a sensible time scale might be the number of divisions a certain cell has undergone; or in engineering the number of items a machine has produced might be a more interesting measure of age than its actual lifetime in months or years. Objectives of survival analysis are to understand how a disease functions, to identify covariates that significantly affect failure times, to study if and how treatments increase life expectancy etc.

We will now introduce very briefly the basic notions that will be used extensively throughout this text. For a detailed discussion of survival analysis see one of the many excellent textbooks that exist on this subject as for example Kalbfleisch and Prentice (2002) and Lawless (2003).

Continuous Failure Times

Let us first consider the case where the random variable T is continuous with cumulative distribution function $F(t)$ and density $f(t)$. In order to characterize the distribution of T, we will use instead of F rather the survivor function,

$$S(t) = \mathrm{P}(T > t) = 1 - F(t)\,,$$

or the hazard function (also called hazard rate, failure rate, or force of mortality)

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(t \le T < t + \Delta t \,|\, T \ge t) \\ &= \frac{f(t)}{S(t)} \\ &= -\frac{\mathrm{d}}{\mathrm{d}t} \log(S(t)) \,. \end{aligned}$$

Let $H(t) = \int_0^t h(u)\,\mathrm{d}u$ be the cumulative hazard. Then we have the relation

$$S(t) = e^{-H(t)} \,,$$

which shows that the hazard function characterizes uniquely the distribution of T. Still another way to give the law of a survival time variable is the expected residual life at t, defined as

$$r(t) = \mathrm{E}(T - t \,|\, T \ge t) \,,$$

which is less general in the sense that it is defined only if T has a finite mean. But if this is the case, $r(t)$ specifies a unique probability distribution.

The hazard function can be interpreted as instantaneous rate of failure at time t, given the individual is still alive at time t. This concept has been extremely successful in survival analysis and $h(t)$ is used extensively. Nonetheless, $h(t)$ is a complicated function and its interpretation is more difficult than one might expect at first sight. For example when a population is observed, the monotonicity properties of the overall hazard can be drastically different from the monotonicity properties of the hazard functions of the single individuals due to a frailty effect. Strictly increasing hazard functions can give rise to a strictly decreasing population hazard under mixing. We will consider this phenomenon more closely in Chapter 3.

Another point that needs some care is the interpretation of the failure rate in terms of probabilities. Neither $h(t)$ nor $\int_a^b h(u)\,\mathrm{d}u$ is a

probability, but $h(t)\,\mathrm{d}t$ is. However, as long as the cumulative hazard is small enough,

$$H(t) \approx 1 - e^{-H(t)} = 1 - S(t)\,.$$

This means $H(t)$ is a good approximation of the probability to fail in the interval $[0, t]$ for rare events or short time periods.

Discrete Failure Times

Analogous definitions can be given if T is a discrete random variable. Let $a_1 < a_2 < \ldots$ be the (nonnegative) time-points with positive probability mass, and let us note

$$f(a_i) = f_i = \mathrm{P}(T = a_i)\,,\ i = 1, 2, \ldots\,.$$

Then, the survivor function is given by the step function

$$S(t) = \sum_{i|a_i > t} f_i\,.$$

In the discrete case, the hazard function at time a_i is a probability, namely the probability of failure at a_i given survival up to a_i,

$$h_i = \mathrm{P}(T = a_i \,|\, T \geq a_i) = \frac{f(a_i)}{S(a_i-)}\,.$$

Consequently, $1 - h_i$ is the probability not to fail at a_i, given survival up to this point. The survivor function can therefore be reconstructed from the hazard in form of a product,

$$S(t) = \prod_{i|a_i \leq t} (1 - h_i)\,.$$

In other words, the survivor function may be seen as arising from a sequence of coin flipping experiments, such that at a_i an item fails with probability h_i and moves on to the next point with probability $1 - h_i$. The cumulative hazard can now be defined in two ways. Either we set

$$H(t) = \sum_{i|a_i \leq t} h_i\,,$$

or we define

$$H^*(t) = -\sum_{i|a_i \leq t} \log(1 - h_i)\,,$$

which has the advantage to keep the relationship $S(t) = \exp(-H^*(t))$. Note that as long as h_i is small, we have $\log(1 - h_i) \approx -h_i$ and thus

$$H(t) \approx H^*(t)\,.$$

In this work, we will take the former definition for the cumulative hazard in models with discrete time.

Besides purely continuous and purely discrete survival times, we can also imagine mixed situations, where T has a continuous and a discrete part. In this case, the cumulative hazard is given by a linear combination,

$$H(t) = \int_0^t h^c(u)\,\mathrm{d}u + \sum_{a_j \leq t} h_j^d\,,$$

and the survivor function can be decomposed in an analogous manner as the product of a continuous and a discrete part.

Discretization of a Continuous Model

Later on in this thesis, we will apply continuous multistage carcinogenesis models to human cancer incidence data. Such data come commonly in the form of counts of cases as a function of age, where ages are binned into groups. That is we do not observe the actual failure time of an item, but we know only during which interval of a given partition $\{[a_{i-1}, a_i),\ i = 1, \ldots, k\}$ the event occurs. If we want to fit a continuous model with survivor function $S(t)$ and hazard function $h(t)$ to such data, we must transform it into the corresponding versions in discrete time. For example the discretized hazard at a_i is

$$\lambda_i = \frac{S(a_i) - S(a_{i+1})}{S(a_i)}\,.$$

So the hazard rate per year at a_i is

$$\begin{aligned}\frac{\lambda_i}{a_{i+1}-a_i} &= \frac{1}{a_{i+1}-a_i}\left(1-e^{-[H(a_{i+1})-H(a_i)]}\right)\\ &\approx \frac{H(a_{i+1})-H(a_i)}{a_{i+1}-a_i}\\ &\approx h(a_i)\,,\end{aligned}$$

where the approximations hold if $H(a_{i+1})-H(a_i)$ and $a_{i+1}-a_i$ are small enough. Note that λ_i is bounded by one since it is a conditional probability. Therefore, also the discretized hazard rate per year is bounded, while the hazard function of the continuous model , $h(t)$, is not. This effect can cause considerable bias. Figure 2 below shows the effect of discretization of a Weibull model as the time axes is split into five year intervals.

For rare events like cancer, where the incidence rates are extremely low, we can directly compare continuous hazard curves to discretized data. If, however, the above approximations are poor, then we must transform the model into its discretized version in order to compare models to data.

Censoring and Estimation

As described above, we will work mainly with grouped data. This is a special form of censoring. In general, censoring means that we do not observe the actual failure time, but only some derived information. In order to illustrate this, let there be given n individuals with lifetimes $T_1,\ldots,T_n$. In the case of grouped data, we do not observe the T_j. Instead we only know into which interval from a given partition $\{[a_{i-1},a_i)\,,\ i=1,\ldots,k\}$ these times fall. A more general scheme would be to replace the fixed intervals by random ones. In this case, we observe instead of T_j a pair of random variables (L_j,U_j), such that $T_j\in[L_j,U_j]$. Another very common situation is right-censoring, which means that for some individuals we only know that their survival time T_j exceeds a possibly random censoring time C_j. Instead of T_j, we

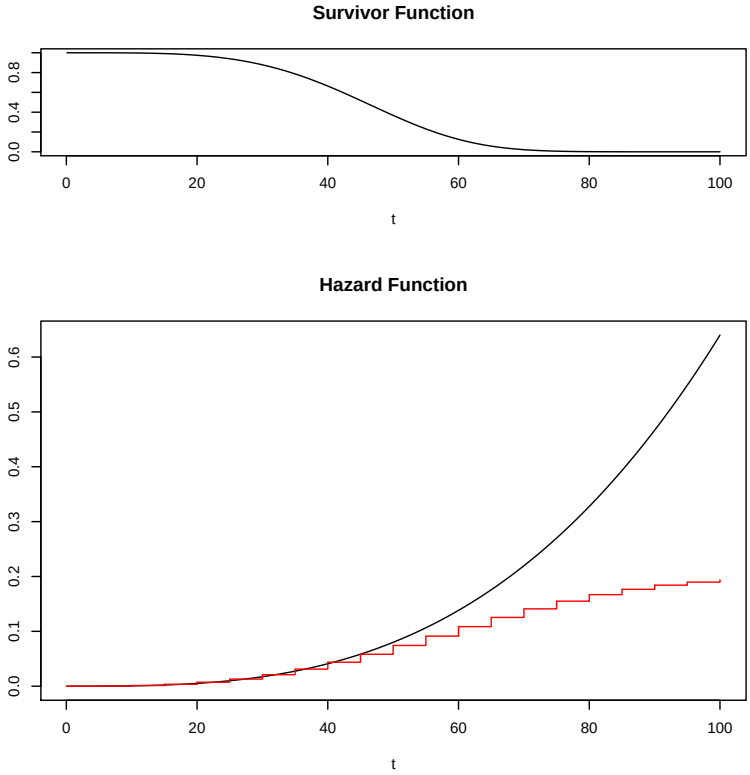

Figure 2: *Survivor and hazard function of the Weibull model with shape parameter* $n = 4$ *and scale parameter* $\theta = 50$ *. The red curve gives the hazard rates per year* $\lambda_i/(a_{i+1} - a_i)$ *, where* $a_i = 5 \cdot i$ *,* $i = 0, 1, \ldots$ *.*

observe $(\min\{T_j, C_j\},\ \mathcal{I}_{\{T_j \le C_j\}})$, where $\mathcal{I}_A$ is the indicator function of event A. Many other censoring-mechanisms are possible and form the core of a lot of issues in survival analysis.

Finally, let us introduce two very useful classical non-parametric estimators - one of the cumulative hazard, and one of the survivor function respectively. For this purpose, let there be given a sample of right censored survival times, (t_i, δ_i), $i = 1, \ldots, n$, where $\delta_i = 1$ if t_i is an observed failure time and $\delta_i = 0$ if t_i is a censoring time. So $\tilde{k} = \sum_{i=1}^{n} \delta_i$ elements out of the list $t_1, \ldots, t_n$ are effectively observed failures. We will denote the time points at which these failures occurred by $a_1 < a_2 < \cdots < a_k$. Note that $k \le \tilde{k}$, and $k < \tilde{k}$ if there are ties. Suppose that d_i items fail at a_i and m_i are censored during the interval $[a_i, a_{i+1})$. Let $n_i = \sum_{j=i}^{k}(d_j + m_j)$ be the population at risk at a_i. A natural way to estimate the discrete hazard at a_i is to take the empirical hazard, that is, the observed failures among the population at risk. Formally,

$$\hat{\lambda}_i = \frac{d_i}{n_i},$$

which leads to the *Nelson-Aalen* estimator for the cumulative hazard,

$$\hat{\Lambda}(t) = \sum_{a_i \le t} \frac{d_i}{n_i}.$$

The function $\hat{\Lambda}(t)$ characterizes a discrete distribution, which has the survivor function

$$\hat{S}(t) = \prod_{a_i \le t} (1 - \hat{\lambda}_i).$$

This estimate $\hat{S}(t)$ is well known in survival analysis. It is the so-called *Kaplan-Meier* or *product limit* estimate of the survivor function.

In order to get variance estimates of $\hat{\Lambda}(t)$ and $\hat{S}(t)$, one exploits the fact that the empirical hazards can be derived using maximum likelihood theory. In fact, it can be shown that the $\hat{\lambda}_i$, $i = 1, \ldots, k$, are asymptotically independent with variances estimated by $\hat{\lambda}_i(1 - \hat{\lambda}_i)/n_i$. This leads to an estimate of the asymptotic variance of the Nelson-

Aalen estimator,

$$\widehat{\mathrm{Var}}(\hat{\Lambda}(t)) = \sum_{a_i \le t} \frac{d_i(n_i - d_i)}{n_i^3} .$$

The above result in combination with the delta method gives an estimation of the variance of the Kaplan-Meier estimator,

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{a_i \le t} \frac{d_i}{n_i(n_i - d_i)} ,$$

called Greenwood's formula.

In the case of grouped data, we may want to slightly modify the above expressions. Let $I_1, \ldots, I_k$ denote the intervals $I_i = [a_i, a_{i+1})$. We can now interpret d_i, m_i, and n_i as the number of failures during I_i, the number of censorings during I_i, and the number of individuals at risk at the beginning of I_i. The empirical hazard as introduced before, $\hat{\lambda}_i = d_i/n_i$, does not account for the fact that also some of the censored individuals may fail before a_{i+1}. Or, in other words, not all the n_i individuals are at risk for the whole interval I_i. A commonly used adjustment replaces $\hat{\lambda}_i$ by

$$\hat{\lambda}'_i = \frac{d_i}{n_i - m_i/2} ,$$

and we get the so-called *life-table estimator* for the survivor function,

$$\hat{S}_{\text{LT}}(a_i) = \prod_{j=1}^{i} (1 - \hat{\lambda}'_j) .$$

We obtain a similar expression for the variance of $\hat{S}_{\text{LT}}(a_i)$ as before by substituting n_i with $n'_i = n_i - m_i/2$ in Greenwood's formula.

1.3.2 Counting Processes

We will now briefly introduce the counting process framework of lifetime models. This formalism is widely used in survival analysis and provides a large supply of concepts and results from probability theory. The

use of counting process notations for statistical analysis of failure time data has been introduced by Aalen (1975). A thorough discussion of the subject give Andersen et al. (1993).

A counting process $\{N(t),\ t \geq 0\}$ is a stochastic process such that $N(0) = 0$ and $N(t)$ counts the number of events that have occurred in the interval $(0, t]$. In the case of a survival study, it is very natural to consider processes that count the number of individuals that have failed up to time t. Let us suppose that the survival times are right censored, that is we observe a sequence of pairs $(\min\{T_i, X_i\},\ \mathcal{I}_{\{T_i \leq X_i\}})$, $i = 1, \ldots, n$. We can now define the underlying counting process

$$\tilde{N}_i(t) = \mathcal{I}_{\{T_i \leq t\}},$$

which counts the events that have occurred during the interval $(0, t]$ for individual i. Similarly, the process of observed failures is

$$N_i(t) = \mathcal{I}_{\{T_i \leq t,\ T_i \leq X_i\}},$$

and the at risk process is

$$Y_i(t) = \mathcal{I}_{\{T_i \geq t,\ X_i \geq t\}}.$$

We can now express the process of observed failures in terms of $\tilde{N}_i(t)$ and $Y_i(t)$,

$$N_i(t) = \int_0^t Y_i(u)\, \mathrm{d}\tilde{N}_i(u).$$

The total number of observed failures is $N_\cdot(t) = \sum_{i=1}^n N_i(t)$, and the population at risk is $Y_\cdot(t) = \sum_{i=1}^n Y_i(t)$.

Throughout this text, we will consider independent censoring, that is T_i and X_i are independent. We will focus on the special case, where X_i are i.i.d. with an unknown distribution function G, and also T_i are supposed to be i.i.d. Their distribution is specified by the cumulative intensity function $\Lambda(t)$, which can now be characterized by

$$\mathrm{d}\Lambda(t) = \mathrm{P}(\,\mathrm{d}\tilde{N}_i(t) = 1 \,|\, \tilde{N}_i(t^-) = 0).$$

In order to link $\Lambda(t)$ to the observed events process $N_i(t)$, we must specify the history,

$$\mathcal{F}_t = \sigma\{N_i(u),\ Y_i(u^+),\ i = 1, \ldots, n;\ 0 \le u \le t\}.$$

With this definition the at-risk processes $Y_i(t)$ are predictable with respect to the filtration $\mathcal{F}_t$. This means that if we know everything that happened up to time t^-, then we also know which individuals are at risk at t and which are not. The assumption of independent censoring leads to the multiplicative intensity model,

$$\mathrm{P}(\,\mathrm{d}N_i(t) = 1 \,|\, \mathcal{F}_{t^-}) = Y_i(t)\,\mathrm{d}\Lambda(t).$$

In our case, $\Lambda(t)$ is a deterministic function, and therefore $Y_i(t)\,\mathrm{d}\Lambda(t)$ is predictable. The power of the counting process framework for deriving asymptotic properties comes from the fact that

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\,\mathrm{d}\Lambda(u)$$

is a zero mean martingale. In other words, $N_i(t)$ splits into two parts: a systematic part, $\int_0^t Y_i(u)\,\mathrm{d}\Lambda(u)$, called the counting process compensator, and a purely random part, $M_i(t)$, which is called the counting process martingale. The martingale $M_i(t)$ is called square integrable if $\mathrm{E}(M_i^2(t)) < \infty$ for all t. If $M_i(t)$ is square integrable, then its predictable variation process is defined as

$$\langle M_i \rangle(t) = \int_0^t \mathrm{Var}(\,\mathrm{d}M_i(u) \,|\, \mathcal{F}_{u^-}).$$

It can be shown, that $\langle M_i \rangle(t)$ is the compensator of $M_i^2(t)$, which implies that $M_i^2(t) - \langle M_i \rangle(t)$ is a mean zero martingale. It follows that

$$\mathrm{Var}(M_i(t)) = \mathrm{E}\big(\langle M_i \rangle(t)\big).$$

In other terms, $\langle M_i \rangle(t)$ is an unbiased estimator of the variance of $M_i(t)$. Similarly, for two individuals we get martingales $M_i(t),\ M_j(t)$,

which are defined on the same filtration. Their predictable covariation process is defined as

$$\langle M_i, M_j \rangle(t) = \int_0^t \mathrm{Cov}(\, \mathrm{d}M_i(u),\, \mathrm{d}M_j(u) \,|\, \mathcal{F}_{u^-}) \,,$$

and $M_i(t),\ M_j(t)$ are said to be orthogonal if $\langle M_i, M_j \rangle(t) = 0$ for all t. Based in $M_i(t)$, we can construct new martingales in quite general ways. Let $G(t)$ be a bounded predictable process, then also

$$U(t) = \int_0^t G(u)\, \mathrm{d}M_i(u)$$

is a mean zero martingale. If $M_i(t)$ is square integrable, then also $U(t)$ will be so, and we have

$$\langle U \rangle(t) = \int_0^t G^2(u)\, \mathrm{d}\langle M_i \rangle(u) \,.$$

The estimators introduced in Section 1.3.1 can be expressed in such a form. Let $J(t) = \mathcal{I}_{\{Y_\bullet(t)>0\}}$ and let us take the convention $0/0 = 0$. Then, for example the Nelson-Aalen estimator becomes

$$\hat{\Lambda}(t) = \int_0^t \frac{J(u)}{Y_\bullet(u)}\, \mathrm{d}N_\bullet(u) \,.$$

Asymptotic properties of such estimators can now be derived using martingale central limit theory. In particular, let us mention the following version of Rebolledo's theorem. Let $U^{(n)}(t) = (U_1^{(n)}(t), \ldots, U_g^{(n)}(t))$ be a vector-valued martingale, where n denotes the size of the population. Suppose that $U_i^{(n)}(t)$ can be written in the form

$$U_i^{(n)}(t) = \int_0^t G_i^{(n)}(u)\, \mathrm{d}M^{(n)}(u) \,,$$

where $G_i^{(n)}(t)$ is a predictable process. We must introduce for given $\epsilon > 0$ the process

$$U_{\epsilon i}^{(n)}(t) = \int_0^t G_i^{(n)}(u) \mathcal{I}_{\{|G_i^{(n)}(u)|>\epsilon\}}\, \mathrm{d}M^{(n)}(u) \,.$$

Let there exist a positive semi-definite $g \times g$-matrix $V(t)$ such that $V(0) = 0$, $V(t) - V(s)$ is positive semi-definite for all $0 \leq s \leq t$, and $V(t)$ right continuous. Then, from

$$\begin{aligned} \langle U^{(n)} \rangle(t) &\xrightarrow{\mathrm{P}} V(t) \quad \text{as } n \to \infty, \text{ and} \\ \langle U_{\epsilon i}^{(n)} \rangle(t) &\xrightarrow{\mathrm{P}} 0 \quad \text{as } n \to \infty \text{ for all } i \text{ and } \epsilon > 0, \end{aligned}$$

it follows that

$$U^{(n)}(t) \xrightarrow{\mathcal{D}} \mathcal{N}(0, V(t)) \quad \text{as } n \to \infty.$$

1.3.3 Mixture Models

One objective of this work is to study multistage carcinogenesis models in relation to heterogeneity. It is well known that susceptibility to cancer varies among individuals due to both genetic and environmental factors. Mixture models will present a way to account for such hidden differences.

Let us consider a continuous failure time T with survivor function $S(t|\theta)$ depending on some parameter θ. If the observed population is not homogeneous with respect to θ, we may suppose that there exists a distribution function $G(\theta)$, which quantifies the variation of θ. The observed population survivor function is

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

A similar expression can be derived for the hazard function of such a

model,

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathrm{P}(t \le T < t + \Delta t \,|\, T \ge t) \\ &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t S(t)} \int \mathrm{P}(t \le T < t + \Delta t \,|\, \theta) \, \mathrm{d}G(\theta) \\ &\overset{(*)}{=} \frac{1}{S(t)} \int f(t|\theta) \, \mathrm{d}G(\theta) \\ &= \int h(t|\theta) \frac{S(t|\theta)}{S(t)} \, \mathrm{d}G(\theta) \,, \end{aligned}$$

where $(*)$ holds if the limit operation and the integration can be interchanged. In the sequel, we will call $S(t)$ a mixture model with mixing distribution $G(\theta)$ and components $S(t|\theta)$.

An important point in building sound models is identifiability. Let $\mathcal{G}$ be a given set of candidate distributions for G. Then, $\mathcal{G}$ induces a set of mixture models,

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

The family of mixture models $\mathcal{S}$ is said to be identifiable (with respect to $\mathcal{G}$), if the implication

$$\int S(t|\theta) \, \mathrm{d}G_1(\theta) \equiv_t \int S(t|\theta) \, \mathrm{d}G_2(\theta) \implies G_1 \equiv_\theta G_2$$

holds for all $G_1, G_2 \in \mathcal{G}$. In other words, the population survivor function must determine uniquely the underlying mixing distribution within a pre-specified family. This condition turns out to be hard to verify in general settings, and useful results exist only for special cases.

1.3.4 Analytic Graduation

All the statistical tools we have described up to now aim at the specification of models for survival data. As a last topic in this short review

of some methodology we will focus on an issue related to parameter estimation. In Chapter 5 we will fit stochastic carcinogenesis models to grouped cancer incidence data. That means we want to estimate a parameter θ from data of the form (r_i, o_i), $i = 1, \ldots, N$, where r_i counts the population at risk at t_i, and o_i counts the observed cases during time interval $[t_i, t_{i+1})$. The basic idea of analytic graduation is to derive raw hazard estimates, $\hat{\lambda}_i$, and then to fit the parametric hazard $h(t;\theta)$ evaluated at some representative time points. For example, one can take the occurence/exposure rates as raw estimates,

$$\hat{\lambda}_i = \frac{o_i}{r_i(t_{i+1} - t_i)},$$

and the midpoints $t_i^* = (t_i + t_{i+1})/2$, where we suppose that t_{N+1} is finite. Let $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_N)^T$ and $h(t^*;\theta) = (h(t_1^*;\theta), \ldots, h(t_N^*;\theta))^T$. Using this notation, we obtain an estimate $\hat{\theta}$ by minimizing the quadratic form

$$Q(\theta) = (\hat{\lambda} - h(t^*;\theta))^T M (\hat{\lambda} - h(t^*;\theta)),$$

where M is a positive definite symmetric matrix. So for example if M is the identity matrix, then $\hat{\theta}$ is the least squares estimate. In order to get asymptotic properties of $\hat{\theta}$, we need regularity conditions. Namely, we make the assumptions

1) θ varies in an open subset Θ of $\mathbb{R}^p$;

2) $J(\theta)$, where $(J(\theta))_{ij} = \frac{\partial}{\partial \theta_j} h(t_i^*;\theta)$, is of full rank for all $\theta \in \Theta$;

3) as a function of θ, $h(t;\theta)$ is one-to-one, bicontinuous, and continuously differentiable;

4) if n is the total number of items in the study, then

$$\sqrt{n}(\hat{\lambda} - \lambda_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_0)$$

as $n \to \infty$ for a vector λ_0 and a positive definite matrix Σ_0.

Let θ_0 denote the true value of θ, then under 1) - 4) we have that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

where

$$\Sigma = (J_0^T M J_0)^{-1} J_0^T M \Sigma_0 M J_0 (J_0^T M J_0)^{-1}$$

and $J_0 = J(\theta_0)$.

The method of analytic graduation is important in demography for the analysis of vital rates. The main motivation for its development was that raw hazard estimates, extracted for example from mortality tables, gave rise to rather rough curves when plotted against age. However, there was supposed to be an underlying smooth mortality curve. Graduation was used to fit such smooth models to the wiggling raw hazard estimates. A detailed discussion can be found in Hoem (1972, 1976).

1.4 Outline

We finish this introduction with a short outline of the remainder of this work. In Chapter 2, we first review the general stochastic multistage carcinogenesis model. Then, we explain the special case that will be used extensively in all subsequent parts of this text. The chapter is very short, since we take up an existing and well known carcinogenesis model.

The aim of this thesis is to introduce the concept of population heterogeneity into the multistage carcinogenesis model. We achieve this by the means of mixture modeling, and our extension to the multistage model is given in Chapter 3.

Chapter 4 treats identifiability of both the multistage model itself and the mixture structure we introduce. We extend known results from the two-stage case to the multistage version of the model. Next, we prove the identifiability of our mixture model.

In Chapter 5, we present applications of the mixture model to human cancer incidence data. In particular, we discuss problems encountered with maximum likelihood estimation and apply analytic graduation, which was described in the previous section, to our case.

Finally, in Chapter 6 we consider population heterogeneity itself and discuss related statistical issues. By doing so, we end this thesis

with a presentation of some of the present challenges for the application of statistical methods in cancer research.

CHAPTER 2

Stochastic Carcinogenesis Models

2.1 Historical Review

The development of mathematical carcinogenesis models began around the 1950s. It was related to the insight that mutations could eventually cause cancer. The corresponding biomedical theory that established the possibility of a mutational origin of tumors can be traced back to the first half of the 20th century and is associated with names such as Karl Heinrich Bauer and Theodor Boveri. See Edler and Kopp-Schneider (2005) for a short review.

One of the first attempts to build a biologically based quantitative description of the process leading to cancer was Nordling (1953), who investigated cancer mortality as a function of age. Using data from several countries, he observed a linear dependency on a log-log scale between mortality λ and age t. He concluded, that $\lambda(t) \propto t^{n-1}$, where n is the number of mutations needed to cause cancer. The data suggested seven mutational events. Armitage and Doll (1954) further developed

this idea. They assumed that successive discrete events are required for tumor development, where the cellular changes involved must not necessarily be mutations. They also studied time varying carcinogenic factors. However, such multi-hit models did not account for the growth dynamics of pre-neoplastic cells. A first two stage theory was proposed by Armitage and Doll (1957) , who used exponential growth of intermediate cells. Kendall (1960) introduced the theory of stochastic processes into carcinogenesis modeling, taking birth-and-death processes to describe cell proliferation.

These early works form the bases of the mathematical multistage theory of carcinogenesis and led finally to the classical two stage model in Moolgavkar and Venzon (1979); Moolgavkar and Knudson (1981). Their *two-stage clonal expansion (TSCE)* model stipulates that normal dividing cells degenerate to cancer cells via two stages: an initiating stage, where a mutation leads to some growth advantage, and a subsequent promotion stage, in which the population of initiated cells expands and thereby increases the number of target cells that might turn into a tumor. The approach stressed the importance of both mutations and clonal expansion in the process leading to cancer. This model is probably the most widely accepted mechanistic model of carcinogenesis. In the literature it is often called *Moolgavkar-Venzon-Knudson (MVK)* two-stage model instead of TSCE model.

The TSCE model has found many applications. Let us mention some of the extensions that have been introduced: Moolgavkar and Luebeck (1990) discuss time dependent parameters, Kopp-Schneider and Portier (1994) incorporate stem cells into the model, Little (1995) investigates the number of mutations needed to cause cancer, and Little and Wright (2003) take genomic instability into account. General multistage models, as the one shown in Figure 3, take up the same structure as the TSCE model, but allow for more than one intermediate stage.

As we will see later on, the TSCE model leads to rather cumbersome expressions, and an exact formula of the survivor function that did not require numerical integration was derived only in Kopp-Schneider et al. (1994) and in Zheng (1994). Another important issue is identifiability,

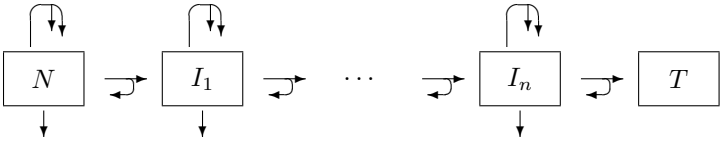

Figure 3: *A general multistage model: normal cells at risk, N, are transformed via $n+1$ discrete events into tumor cells, T. The events are supposed to be irreversible, and the normal and intermediate cells divide and die or differentiate. Division is either symmetric and leads to two identical daughter cells, or it is asymmetric and leads to a normal daughter cell and a daughter cell entering the next compartment. Growth of the tumor cells is not considered.*

which will be discussed extensively in Chapter 4.

Due to this long history, we should not have in mind a single model when talking about the multistage model. We should rather think of 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 Whittemore and Keller (1978), Tan (1991), Kopp-Schneider (1997) and Edler and Kitsos (2005).

2.2 The Multi-Hit Model

The multi-hit model of carcinogenesis assumes that a normal cell is transformed into a cancer cell via a sequence of n discrete events, or hits, as shown in Figure 4.

The model does not account for cell dynamics such as cell division and differentiation. The structure given in Figure 4 is the one proposed in the articles of the early 1950s. Armitage and Doll (1954)[1] show that

[1]Note that in their article, Armitage and Doll talk of a *multi-stage theory of carcinogenesis*. Nevertheless, we will use the term *multi-stage* only in connection with clonal expansion models as the one given in the next section.

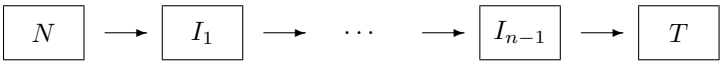

Figure 4: *A normal cell is transformed into a cancer cell via discrete events that occur at rates* $\nu_1, \ldots, \nu_n$.

the hazard rate for the transformation of a single cell via n steps is approximately

$$h(t) = \frac{\nu_1 \ldots \nu_n}{(n-1)!} t^{n-1} \tag{1}$$

if the hits must occur in a given order. If we assume that the events can happen in any order, we must multiply the above expression by $n!$. This process is equally likely to take place in any of the cells at risk. If we suppose the number of cells at risk, N_0, to be constant over time, then the overall incidence rate for an organism is N_0 times the incidence rate for a single cell.

Note that Armitage and Doll used an approximation to derive (1). Let us consider a single hit that can occur at any moment. Then, the time to event is exponentially distributed, $X \sim \mathcal{E}(\nu)$, and the probability of the event to happen in the interval $(0, t)$ is

$$\mathrm{P}(X \in (0,t)) = 1 - e^{-\nu t} = \nu t + o(\nu t)\,.$$

The hazard function of the multi-hit model with n hits is derived in the following way. The first $n-1$ hits occur in $(0, t)$. So under independence and for equal rates the corresponding probability is approximately $(\nu t)^{n-1}$. If we impose a given order, we must divide this probability by $(n-1)!$. The last event happens in the short interval $(t, t + \mathrm{d}t)$, with probability $\nu\,\mathrm{d}t$. Putting these parts together, we get the probability of the nth transformation to happen at time t, namely $h(t)\,\mathrm{d}t$, where $h(t)$ is as given in (1).

This approximation is valid as long as νt is small. Moolgavkar (2004) and Moolgavkar and Luebeck (2003) state that this simplified model can be applied if the probability of getting cancer is low - as usually is the case in human population data. But the approximation is poor in other settings like experiment data with high tumor probability. In any case, the theoretical properties as for example the asymptotic behaviour change with respect to the exact solution.

From now on, we will call *multi-hit model* the version given by the hazard function

$$h_{\text{MH}}(t) = n\nu^n N_0 t^{n-1}\,.$$

That is, we only consider the case with equal rates $\nu_1 = \cdots = \nu_n = \nu$ for the n hits. The corresponding survivor function is

$$S_{\text{MH}}(t) = e^{-\nu^n N_0 t^n}\,.$$

In other words, the waiting time for tumor onset is modelled as a Weibull with shape parameter n and scale parameter $(\nu \sqrt[n]{N_0})^{-1}$.

2.3 The Multistage Carcinogenesis Model

The structure of a general multistage model has been given in Figure 3. The classical approach to derive the survivor function of such a model uses the probability generating function,

$$\varphi(s;t) = \sum_{i_0,\ldots,i_{n+1}} s_0^{i_0} \cdot \ldots \cdot s_{n+1}^{i_{n+1}} \cdot \mathrm{P}(N(t) = i_0, \ldots, T(t) = i_{n+1})\,,$$

of the vector

$$(N(t), I_1(t), \ldots, I_n(t), T(t))\,,$$

which counts the number of cells in the different compartments. Kolmogorov's equations yield differential equations for $\varphi(s;t)$, which may eventually be solved along characteristics.

Throughout this text, we will work with a simplified version of the multistage model, but a version that is general enough to incorporate all the main features of the carcinogenesis process. The model is shown

schematically in Figure 5. Note that we get the classical TSCE model if $n = 1$. We see also that the step called initiation in this multistage model corresponds precisely to the multi-hit model, but as an extension the multistage model incorporates also promotion.

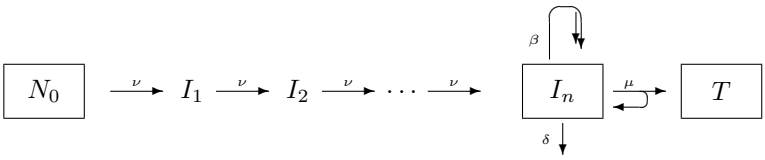

Figure 5: *N_0 denotes the number of normal cells. To get initiated, a normal cell accumulates n consecutive mutations, where ν denotes the mutation rate per cell per year for the gene in question. The number of cells having accumulated k mutations is noted I_k, $1 \leq k \leq n$. The fully initiated cells, I_n, divide into two identical daughter cells and differentiate or die. Eventually, a cell out of a clone of initiated cells may divide asymmetrically and give rise to a tumor cell T. Note that $\mu/(\mu + \beta)$ can be interpreted as the probability that a promoted cell is created during a cell division.*

The same structure has been used by Luebeck and Moolgavkar (2002), who derived the exact survivor and hazard functions for the cases $n = 1, 2, 3, 4$. In our work, we will use an approximation of the Armitage-Doll type for initiation. This means we use the same model as Morgenthaler et al. (2004), who give a detailed discussion and derivation for this case. More precisely, 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 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 (promotion), 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 susceptible cells in the normal state and the different cells act independently of one another.

Under the above conditions, the survivor function for tumor onset can be represented as

$$S(t) = \exp\left\{-\int_0^t \lambda_I(x) F_P(t-x)\,\mathrm{d}x\right\}, \tag{2}$$

where the intensity of initiation is

$$\lambda_I(x) = n\nu^n N_0 x^{n-1}, \tag{3}$$

and $F_P(x)$ is the cdf for the waiting time for the first malignant transformation within a clone starting with one initiated cell at time 0. This function is improper since a clone of initiated cells may die out with a probability greater than 0. Its exact form is

$$F_P(x) = \frac{(\beta-\delta-\mu+\Delta)(\beta-\delta-\mu-\Delta)(e^{-\Delta x}-1)}{2\beta[(\beta-\delta-\mu+\Delta)e^{-\Delta x}-(\beta-\delta-\mu-\Delta)]}, \tag{4}$$

where $\Delta = \sqrt{(\beta+\delta+\mu)^2-4\beta\delta}$. The hazard function can be easily calculated from the survivor function, and we get

$$h(t) = \int_0^t \lambda_I(x) F_P'(t-x)\,\mathrm{d}x,$$

where $(\cdot)' = \frac{\mathrm{d}}{\mathrm{d}t}(\cdot)$. In order to deduce the asymptotic behaviour of the hazard for several n, the following representation is useful. First note that

$$\int_0^t \lambda_I(x) F_P(t-x)\,\mathrm{d}x = \int_0^t \lambda_I(t-x) F_P(x)\,\mathrm{d}x.$$

So we can calculate the hazard also as

$$h(t) = \frac{\mathrm{d}}{\mathrm{d}t} \int_0^t \lambda_I(t-x) F_P(x)\, \mathrm{d}x\,.$$

This implies that

$$h(t;n) = \begin{cases} \nu N_0 F_P(t)\,, & \text{if } n = 1\,; \\ -n\nu \log(S(t;n-1))\,, & \text{if } n \geq 2\,; \end{cases}$$

or in terms of a recursion,

$$h(t;n) = \begin{cases} \nu N_0 F_P(t)\,, & \text{if } n = 1\,; \\ n\nu \int_0^t h(u;n-1)\, \mathrm{d}u\,, & \text{if } n \geq 2\,. \end{cases}$$

This shows that for $n = 1$ the hazard levels off as t goes to infinity. More precisely, the hazard of the TSCE model goes to the finite asymptote

$$\nu N_0 \cdot \mathrm{P}(\text{a clone of initiated cells survives})\,.$$

But the hazard grows to infinity if $n \geq 2$. In any case, $h(t;n)$ is strictly monotonic increasing with t.

CHAPTER 3

Mixtures of Multistage Models

3.1 Heterogeneity

In Chapter 2 we presented models that describe carcinogenesis at a cellular level. We then assumed independent and identical behavior of all the cells at risk of an organism to get a model for a single individual. If we apply this model directly to human cancer incidence data, then the same model is used for the whole population. However, the risk of developing cancer is not equal for everyone. One can easily think of many biological mechanisms and environmental settings that would induce heterogeneity. For example germ line mutations in oncogenes or tumor suppressor genes produce inheritable factors. This leads to newborns that start life with all cells in some intermediate stage. But also a genetically based overall predisposition to cancer for some individuals is plausible. Moreover, exposure to carcinogens varies due to changes in environment and occupation. Therefore, incidence rates are specific for different population subgroups. Such considerations are common is-

sues in survival analysis. For example Vaupel and Yashin (1985) give a very nice introduction into the effects of unobserved heterogeneity, and Aalen (1994) and Hougaard (1995) review frailty modeling in survival analysis.

In our work, we aim at introducing population heterogeneity into the multistage carcinogenesis model presented in Section 2.3. First of all, let us have a brief look at the approaches that have been proposed up to now. Some work has been done by Tan (1988), who developed a general mixed model of one-stage and two-stage models, and Tan and Singh (1990), who applied the special case of a one-stage and a two-stage model to data from retinoblastoma, a cancer of the eye affecting most often small children. Their model is particularly well suited for the description of cancers with early onset for a part of the population, since the one-stage component will cause a peak in the overall hazard at young ages. Biological theory strongly supports their mixture model. Retinoblastoma (RB) is caused by mutations in both copies of a gene called *RB1*, and thus heterozygous individuals have a hereditary predisposition. Additionally, RB does occur in both forms, hereditary and non-hereditary.

Aalen and Tretli (1999) and Moger et al. (2004) use standard frailty modeling to analyze testicular cancer incidence data from Scandinavian registries. More precisely, they model the hazard function for an individual as the product of a non-negative frailty Z and a baseline hazard function, $h_{\text{ind}}(t|Z) = Z \cdot h_0(t)$. They chose the multi-hit hazard as baseline $h_0(t)$ and model Z by the compound Poisson distribution. This frailty distribution allows for two main underlying states for an individual: either $Z = 0$, that is the individual is not susceptible, or $Z > 0$, which means the individual is susceptible and its relative risk with respect to the baseline is distributed as the sum of independent gamma variables. Both states occur with strictly positive probability and the proportion of non-susceptibles, $\mathrm{P}(Z = 0)$, is a parameter of main interest. A more basic version of the same model investigate Izumi and Ohtaki (2004). These authors assume Z be gamma distributed. In other words, they take the Weibull-gamma frailty model and apply it

to data from Japanese atomic bomb survivors.

A different approach is proposed by Morgenthaler et al. (2004), who introduce two new population parameters to describe variability among individuals. The first one, called fraction at risk, quantifies the proportion of susceptibles. The second one, called fraction of deaths due to the cancer among all deaths due to either cancer or related causes, describes the behaviour of competing but related risks. These authors combine the two parameters with the multistage model given in Section 2.3 and analyze lung cancer incidence data from US birth cohorts.

The multistage model is a mechanistic model, so all its parameters have a natural biological interpretation. It is therefore desirable to introduce also the notion of frailty in a biologically meaningful way into the model. The extension by mixture presents such a possibility, and we will take up this approach in our work. We can focus attention on any of the parameters and not just on the number of mutations that are necessary to induce a cancer. If we assume a parameter be a random variable, then the population hazard arises as a mixture. Before we study mixtures of carcinogenesis models, let us make some general remarks on the hazard rates we will get.

3.2 Hazard Rates of Mixture Models

Recall Section 1.3.3, where we have seen that under suitable regularity conditions, the hazard rate of a mixture model can be written as

$$h(t) = \int h(t|\theta) \frac{S(t|\theta)}{S(t)} \, \mathrm{d}G(\theta) \,, \tag{5}$$

where $G(\theta)$ is the mixing distribution function and $h(t|\theta)$, $S(t|\theta)$ are the component hazard and survivor functions. Two features of $h(t)$ will be of main interest for us: monotonicity, and the asymptotic behaviour. Especially, it is important to know if and under which conditions the properties of the components $h(t|\theta)$ are invariant under mixing and still hold for $h(t)$. Regarding monotonicity, we must distinguish the two

cases of distributions with an increasing failure rate (IFR), and distributions with a decreasing failure rate (DFR). From reliability theory it is known that the DFR property is preserved under mixture. However, one can show that the mixture of IFR distributions can be DFR. See Barlow and Proschan (1975), Lynch (1999)and Shaked and Spizzichino (2001) for a detailed account. Gurland and Sethuraman (1995) extend these results to the notion of ultimately IFR/DFR, meaning that the monotonicity property holds for all t larger than some threshold value t_0 . In other words, it can be useful to consider separately the behaviour of $h(t)$ at high ages. General results on the asymptotic properties of $h(t)$ give Block et al. (2003).

In order to investigate the properties of $h(t)$, it is useful to distinguish cases. If the mixing distribution $G(\theta)$ is finite with probability mass π_i at θ_i , then equation (5) becomes

$$h(t) = \pi_1 \frac{S(t|\theta_1)}{S(t)} h(t|\theta_1) + \cdots + \pi_n \frac{S(t|\theta_n)}{S(t)} h(t|\theta_n) \,. \tag{6}$$

That is the weight of component $h(t|\theta_i)$ is time dynamic and equals the proportion of survivors of group i at time t among all survivors at time t . Another useful representation is obtained exploiting the fact that

$$S(t) = \pi_1 S(t|\theta_1) + \cdots + \pi_n S(t|\theta_n) \,.$$

Introducing this expression into (6), we obtain

$$\begin{aligned} h(t) &= \frac{\pi_1 S(t|\theta_1) h(t|\theta_1) + \cdots + \pi_n S(t|\theta_n) h(t|\theta_n)}{\pi_1 S(t|\theta_1) + \cdots + \pi_n S(t|\theta_n)} \\ &= \frac{\pi_1 \frac{S(t|\theta_1)}{S(t|\theta_n)} h(t|\theta_1) + \cdots + \pi_{n-1} \frac{S(t|\theta_{n-1})}{S(t|\theta_n)} h(t|\theta_{n-1}) + \pi_n h(t|\theta_n)}{\pi_1 \frac{S(t|\theta_1)}{S(t|\theta_n)} + \cdots + \pi_{n-1} \frac{S(t|\theta_{n-1})}{S(t|\theta_n)} + \pi_n} \,. \end{aligned}$$

If we have an ordering of the survivor functions such that

$$\lim_{t \to \infty} \frac{S(t|\theta_i)}{S(t|\theta_n)} = 0 \,, \ \forall \, i < n \,, \tag{7}$$

then this implies that asymptotically $h(t)$ will behave like the hazard of the strongest component of the mixture, $h(t|\theta_n)$. Condition (7) means

that the dying out of the nth group comes later than the dying out of all other groups. We will see this condition again in relation with identifiability in the next chapter.

In the continuous case, the population hazard

$$h(t) = \int h(t|\theta) \frac{S(t|\theta)}{S(t)} g(\theta)\, \mathrm{d}\theta$$

is often difficult to investigate unless strong restrictions are put on $h(t|\theta)$ and $g(\theta)$.

In conclusion, we must realize that the hazard rate is a difficult concept, especially if we consider a heterogeneous population. Due to a frailty effect, we can observe hazard rates with peaks, while all individual hazards are strictly monotonic; or IFR distributions can become DFR under mixing. This raises the question which underlying processes can generate an observed population hazard curve and which mechanisms would not be able to do so. An approach that is related to this issue is presented in Aalen and Gjessing (2001), who consider first passage time models for survival analysis.

3.3 Mixing the Multi-Hit Model

In a first step, we will consider the multi-hit model presented in Section 2.2. Recall the hazard and survivor function,

$$\begin{aligned} h_{\text{MH}}(t) &= n\nu^n N_0 t^{n-1}\,, \\ S_{\text{MH}}(t) &= e^{-\nu^n N_0 t^n}\,, \end{aligned}$$

where n counts the number of hits, ν is the rate at which such hits occur, and N_0 is the number of susceptible cells. The multi-hit model has the mathematically appealing Weibull form, and it is a natural simplification of the multistage model. Therefore, it is a useful structure for a preliminary study of the extension of carcinogenesis models through mixture. We will consider the multistage model in the next section.

A Finite Mixture Model

Let us consider the number of mutations needed to cause a cancer. If we assume different numbers of mutations for different population subgroups, then we get a finite mixture with population survivor function

$$S(t) = \sum_{i=1}^{k} \pi_i S_{\text{MH}}(t|n_i)\,, \tag{8}$$

where $1 \le n_1 < n_2 < \cdots < n_k$ are non-negative integers, $0 < \pi_i < 1$, and $\pi_1 + \cdots + \pi_k = 1$. Note that we have

$$\frac{S_{\text{MH}}(t|n_i+1)}{S_{\text{MH}}(t|n_i)} = \exp\left\{-\nu^{n_i} N_0 t^{n_i}(\nu t - 1)\right\} \xrightarrow{t \to \infty} 0\,.$$

According to equations (6) and (7), this shows that the population hazard $h(t)$ approaches component $h_{\text{MH}}(t|n_1)$ as t increases. In particular, if $n_1 = 1$, then the population hazard $h(t)$ levels off at the value νN_0 as $t \to \infty$. But if $n_1 > 1$, then $h(t)$ is ultimately increasing to infinity. Whether $h(t)$ is strictly monotonic increasing for all $t \in [0, \infty)$ or has peaks depends on the specific combination of components and weights $(n_i, \pi_i),\ i = 1, \ldots, k$. For many typical cases the population hazard has peaks as the different population subgroups die out.

Figure 6 illustrates the survivor function and the hazard for a population with three groups. A very small one that needs only four mutations to cancer leads to a peak at young ages. A second group, quite small as well, of high risk individuals, which cause an increase in cases from the age of about forty. And a last group that is at very low risk. Groups two and three have a behavior close enough to cause $h(t)$ to be strictly monotonic from about the age of forty; the population hazard has only one peak, which corresponds to the dying out of the first group.

A Continuous Mixture Model

We can also consider continuous mixture models. Biologically, it would be most natural to focus on the rate ν and to model this parameter with

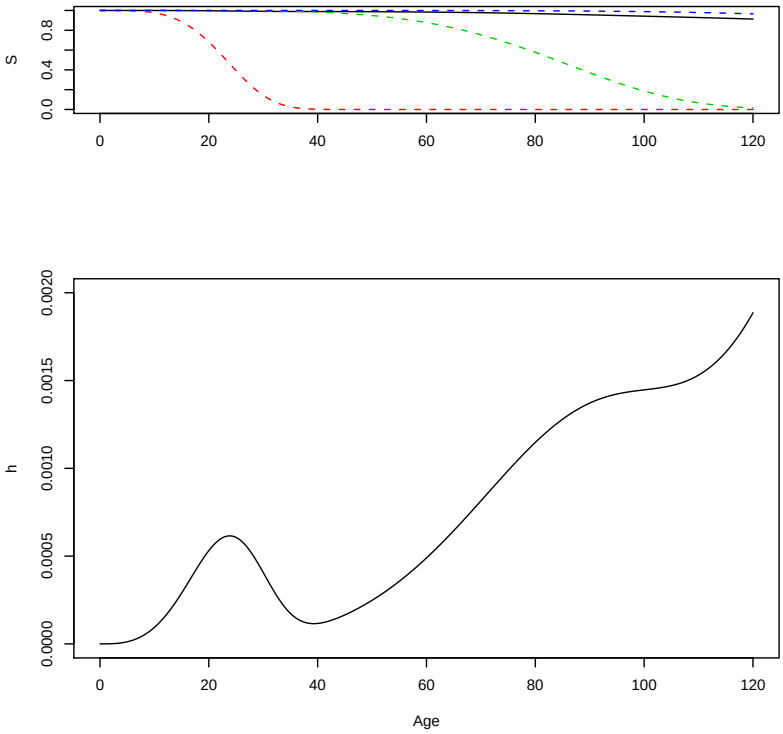

Figure 6: *Population survivor function $S(t)$ and population hazard function $h(t)$ for a multi-hit mixture model. The three components differ by the values chosen for parameter n, namely $n_1 = 4$, $n_2 = 5$ and $n_3 = 6$. The mixing weights are $\pi_1 = 0.01, \pi_2 = 0.045$ and $\pi_3 = 0.945$. The remaining parameters are equal for all components with values $\nu = 7 \cdot 10^{-5}$, $N_0 = 10^{11}$. The dashed lines give the individual survivor functions $S_{MH}(t|n_1)$ (red), $S_{MH}(t|n_2)$ (green) and $S_{MH}(t|n_3)$.*

some continuous distribution. Due to the special structure of the multi-hit model, however, we will make some rather general observations here.

Let us write the survivor function of the multi-hit model in the following form,

$$\tilde{S}_{\text{MH}}(t) = e^{-\theta t^n}\,, \tag{9}$$

where $\theta > 0$, and $n > 1$ is considered fixed. Then, $\tilde{S}_{\text{MH}}(t)$ is a particular case of a survivor function of the form

$$\mathcal{S} = \left\{ S_\lambda(t) = e^{-\lambda G(t)}\,,\ \lambda > 0 \right\}\,,$$

where $G(t)$ is an increasing convex function such that $G(0) = 0$ and $G(t) \uparrow \infty$ as $t \uparrow \infty$. The set $\mathcal{S}$ is called an IFR Lehmann family based on the cumulative hazard function $G(t)$. In other words, $\mathcal{S}$ consists of all survivor functions that stem from a proportional hazards model with baseline cumulative hazard $G(t)$, since the cumulative hazard corresponding to $S_\lambda(t)$ is

$$H_\lambda(t) = \lambda G(t)\,.$$

Gurland and Sethuraman (1995) study mixture models with kernel $\mathcal{S}$ for several mixing distributions for λ. They are mainly interested in cases such that the IFR property of $G(t)$ is reversed, that is, where the hazard of the mixture is either decreasing or ultimately decreasing. In particular, these authors show that if we take the survivor function (9) and model θ with a gamma distribution, then we get a model that is ultimately DFR. Note that their result does not impose conditions on the gamma distribution chosen.

The fact that the pooling of IFR Weibull variables can yield a DFR variable crucially depends on the mixing distribution used. In the case of the multi-hit model with (the aggregated) parameter θ, we cannot take the gamma distribution if we wish to keep a biologically reasonable setting. As a minimal requirement we must assume that the density $f(\theta)$ has bounded support. Let us consider the case

$$\text{supp}\{f\} = [\theta_l, \theta_u] \subset \mathbb{R}_{>0}\,.$$

For the hazard function

$$\tilde{h}_{\text{MH}}(t|\theta) = \theta n t^{n-1}\,,$$

we have the natural ordering $\tilde{h}_{\text{MH}}(t|\theta_1) \leq \tilde{h}_{\text{MH}}(t|\theta_2)$, whenever $\theta_1 \leq \theta_2$. In particular, we have

$$\tilde{h}_{\text{MH}}(t|\theta_l) \leq \tilde{h}_{\text{MH}}(t|\theta) \leq \tilde{h}_{\text{MH}}(t|\theta_u)\,,\ \forall\, t\,,$$

which implies

$$\tilde{h}_{\text{MH}}(t|\theta_l) \leq \tilde{h}(t) = \int \tilde{h}_{\text{MH}}(t|\theta)\frac{\tilde{S}_{\text{MH}}(t|\theta)}{\tilde{S}(t)} f(\theta)\,\mathrm{d}\theta \leq \tilde{h}_{\text{MH}}(t|\theta_u)\,,\ \forall\, t\,.$$

In other words, the population hazard $\tilde{h}(t)$ lies within the area spanned by the strongest component, $\tilde{h}_{\text{MH}}(t|\theta_l)$, and the weakest one, $\tilde{h}_{\text{MH}}(t|\theta_u)$. This implies that $\tilde{h}(t) \to \infty$ as $t \to \infty$. We can resume this result for the general case of Lehmann families.

Proposition 1. *Let $\mathcal{S}$ be an IFR Lehmann family based on hazard function $g(t)$ and suppose that $g(t) \uparrow \infty$ as $t \to \infty$. Let $f(\lambda)$ be a probability density on $(0, \infty)$. Let $h(t)$ be the hazard function of the mixture model with kernel $\mathcal{S}$ and mixing density f. If f is such that* $\text{supp}\{f\} \not\ni 0$*, then $h(t) \to \infty$ as $t \to \infty$.*

Proof: $\mathcal{S}$ defines a proportional hazards model. Therefore $\lambda_1 \leq \lambda_2$ implies that $h_{\lambda_1}(t) \leq h_{\lambda_2}(t)$, $\forall\, t$. Condition $\text{supp}\{f\} \not\ni 0$ implies that the minimum $\lambda_l = \min(\text{supp}\{f\})$ is positive, $\lambda_l > 0$. Since $h_{\lambda_l}(t) \leq h(t)$, $\forall\, t$, the result follows. □

In particular, if $0 \notin \text{supp}\{f\}$, then the population hazard can neither be DFR nor ultimately DFR. In other words, the condition $0 \in \text{supp}\{f\}$ is necessary but not sufficient for a continuous mixture of IFR distributions be DFR. The condition means that the mixture model contains components that are as strong as you like.

As an example, let us consider the mathematically easiest situation of the uniform mixing distribution,

$$\theta \sim \mathcal{U}[\theta_l, \theta_u]\,.$$

In this case it is possible to derive closed form solutions for $\tilde{S}(t)$ and $\tilde{h}(t)$. For $t > 0$, we get the population survivor function,

$$\tilde{S}(t) = \int_{\theta_l}^{\theta_u} e^{-\theta t^n} \frac{\mathrm{d}\theta}{\theta_u - \theta_l} = \frac{1}{(\theta_u - \theta_l)t^n}\left(e^{-\theta_l t^n} - e^{-\theta_u t^n}\right)\,,$$

and the population hazard,

$$\begin{aligned}
\tilde{h}(t) &= \int_{\theta_l}^{\theta_u} \theta n t^{n-1} \frac{e^{-\theta t^n}}{\tilde{S}(t)} \cdot \frac{\mathrm{d}\theta}{(\theta_u - \theta_l)} \\
&= \frac{n t^{2n-1}}{e^{-\theta_l t^n} - e^{-\theta_u t^n}} \int_{\theta_l}^{\theta_u} \theta e^{-\theta t^n} \, \mathrm{d}\theta \\
&= \frac{n}{t} + n t^{n-1} \frac{\theta_l e^{-\theta_l t^n} - \theta_u e^{-\theta_u t^n}}{e^{-\theta_l t^n} - e^{-\theta_u t^n}} \\
&= \frac{n}{t} - \frac{1}{e^{(\theta_u - \theta_l) t^n} - 1} \tilde{h}_{\mathrm{MH}}(t|\theta_u) + \frac{1}{1 - e^{-(\theta_u - \theta_l) t^n}} \tilde{h}_{\mathrm{MH}}(t|\theta_l) \, .
\end{aligned}$$

For $t = 0$, we get $\tilde{S}(t) = 1$, and since we assume $n > 1$, we have

$$\tilde{h}_{\mathrm{MH}}(0|\theta) = 0 \, , \ \forall \, \theta \, ,$$

and therefore $\tilde{h}(0) = 0$. The above expression shows that $\tilde{h}(t)$ approaches $\tilde{h}_{\mathrm{MH}}(t|\theta_l)$ as t gets large. So $\tilde{h}(t)$ is at least ultimately strictly monotonic increasing. Whether $\tilde{h}(t)$ has a peak or not depends on the wideness of the interval $[\theta_l, \theta_u]$. The population hazard will decrease at some ages due to a frailty effect only if θ is allowed to vary enough. This effect is shown in Figure 7.

It would be reasonable to consider also other distributions for θ than the uniform. In such cases, it is much harder to get closed form solutions for the population functions and numerical integration must be used.

3.4 Multistage Carcinogenesis Models

We will now focus attention on the multistage model introduced in Section 2.3. Recall the survivor function of this model,

$$S(t|n, \psi) = \exp\left\{ - \int_0^t \lambda_I(x) F_P(t - x) \, \mathrm{d}x \right\} \, ,$$

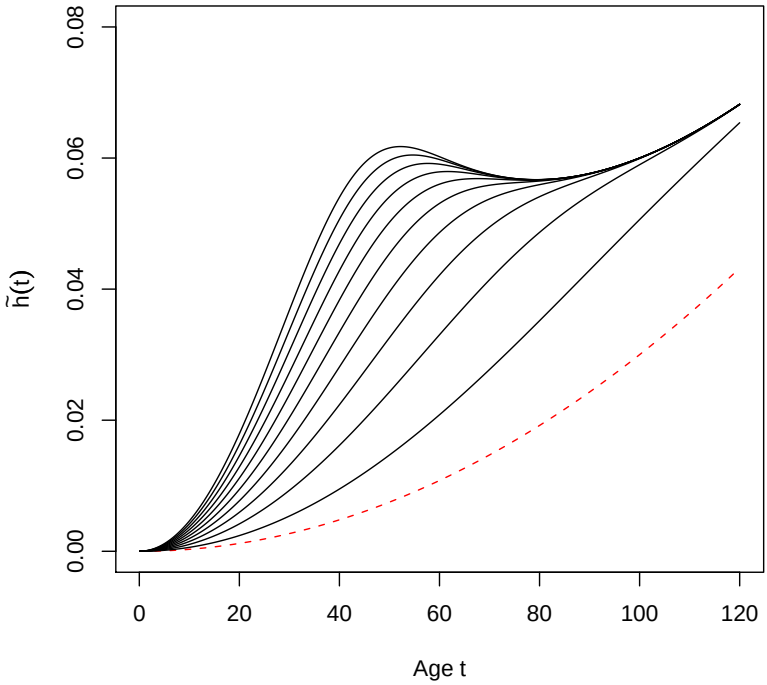

Figure 7: *Population hazard function* $\tilde{h}(t)$ *for the continuous multi-hit mixture model where* $\theta \sim \mathcal{U}[\theta_l, \theta_u]$ *. For all curves, we set* $n = 3$ *and* $\theta_l = 10^{-6}$ *. The upper bound was* $\theta_u = c \cdot \theta_l$ *for* $c = 3, 6, 9, \ldots, 30$ *. The red dashed line corresponds to* $\tilde{h}_{MH}(t|\theta_l)$ *.*

where

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

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

with $\Delta = \sqrt{(\beta+\delta+\mu)^2 - 4\beta\delta}$. In the above expressions, n counts the number of events needed for initiation of a cell, ν is the rate of these events, β and δ are the birth and death rates of the fully initiated cells, and μ is the rate at which cells out of the resulting clone are transformed into cancer cells. These rates are resumed in the vector $\psi = (\nu, \beta, \delta, \mu)$. The remaining parameter N_0 counts the number of cells at risk.

In the sequel, we will introduce heterogeneity through the parameters n and ψ. The motivation behind this is that carcinogenic agents are supposed to have an impact on the biological parameters that control the initiation-promotion scheme. We will consider the biological parameters one at a time. This assumes carcinogens that act as pure initiators or pure promoters. For many known agents this is an oversimplification, since most substances are thought to act upon both stages of carcinogenesis.

For simplicity, we will use interchangeably the notations $S(t|n)$, $S(t|\mu)$ etc. for the multistage model $S(t|n, \psi)$, and $S(t)$ will denote the population survivor function. The same remark applies to the hazard functions. Finally, the number N_0 will be considered fixed for the whole population. This value is usually very large, and even if differences among individuals certainly occur, we suppose them be negligible. If a generalization seems appropriate, then rather to replace the fixed constant N_0 by a time dependent function $N(t)$ as has been proposed by many authors.

Initiation

The multistage model we consider here describes initiation as a sequence of n discrete events, namely rate limiting mutations, which lead to a cell capable of accelerated growth. Similarly to the previous section, it is a natural extension to let n be random. A biological mechanism generating this kind of heterogeneity are germ line mutations of the corresponding genes. This would lead to individuals starting life with all cells in an intermediate stage. Mathematically, this means that

the population survivor function is

$$S(t) = \sum_{i=1}^{k} \pi_i S(t|n_i, \psi)\,, \tag{10}$$

where $0 < \pi_i < 1$ and $\pi_1 + \cdots + \pi_k = 1$. The model determined by (10) shows very much the same behaviour as the finite mixture model of the multi-hit model (8) studied in Section 3.3. Again, the population hazard typically has peaks as the different subgroups die out.

This model can be interpreted alternatively as a multiple pathway model. Suppose that k pathways lead to cancer, where the cascades differ only by the number of rate limiting events. If π_i is the probability a cell enters the cascade with n_i events for initiation, then the survivor function is given by (10). This model is illustrated in Figure 8.

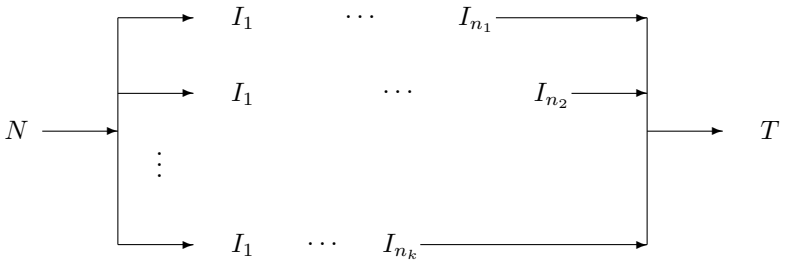

Figure 8: *Normal cells N are transformed into cancer cells T via different cascades involving several stages of degenerated cells I_j. Pathway number i involves n_i mutational events and is taken with probability π_i.*

The second parameter related to initiation is the rate ν, and finite as well as continuous distributions could be used to model its heterogeneity. Also in this case we get an expression resembling the structures studied in the previous section,

$$S(t|\nu) = e^{-\nu^n G(t)}\,,$$

where $G(t)$ is convex, increases to infinity and does not depend on ν. Therefore, we will not go further into details here and turn to the next stage of carcinogenesis.

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 processes carcinogenic agents can influence: the growth of initiated cells, and the malignant transformation of initiated cells.

Growth of initiated cells is modelled via a birth and death process with birth rate β and death or differentiation rate δ. The parameter of real interest, however, is the net growth rate, $\gamma = \beta - \delta$.

The most natural approach is probably to model γ as a continuous random variable with values in a given bounded interval $[\gamma_l, \gamma_u] \subset \mathbb{R}_{>0}$. For example, we can take a beta distribution with density

$$f(\gamma) = \frac{1}{B(a,b) \cdot (\gamma_u - \gamma_l)^{a+b-1}} (\gamma - \gamma_l)^{a-1} (\gamma_u - \gamma)^{b-1} ,$$

for $\gamma \in [\gamma_l, \gamma_u]$. As before, we get the population survivor function $S(t) = \int_{\gamma_l}^{\gamma_u} S(t|\gamma) f(\gamma) \, \mathrm{d}\gamma$. Also the population hazard can be represented in the usual form. With $F(t) = 1 - S(t)$, we get

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \downarrow 0} \int \frac{F(t + \Delta t|\gamma) - F(t|\gamma)}{\Delta t} f(\gamma) \, \mathrm{d}\gamma \\ &\overset{(\star)}{=} \frac{1}{S(t)} \int f(t|\gamma) f(\gamma) \, \mathrm{d}\gamma \\ &= \int h(t|\gamma) \frac{S(t|\gamma)}{S(t)} f(\gamma) \, \mathrm{d}\gamma . \end{aligned}$$

The interchanging of limit and integral signs in step $(\star)$ can be justified

by the dominated convergence theorem. The function

$$F(t|\gamma) = 1 - e^{-\int_0^t \lambda_I(x) F_P(t-x|\gamma)\,\mathrm{d}x}$$

is differentiable and strictly monotonic increasing in t for any $\gamma \in [\gamma_l\,,\,\gamma_u]$. Therefore, the function

$$g(\Delta t, \gamma) = \frac{F(t+\Delta t|\gamma) - F(t|\gamma)}{\Delta t}$$

is continuous on the compact $C_h = [0\,,\,h] \times [\gamma_l\,,\,\gamma_u]$ for any $t, h > 0$. This implies that on C_h the function $g(\Delta t, \gamma)$ takes on its minimum and maximum values,

$$0 < m_h \le g(\Delta t, \gamma) \le M_h < \infty\,,\ \text{for } (\Delta t, \gamma) \in C_h\,,$$

and for example $\int M_h f(\gamma)\,\mathrm{d}\gamma = M_h$. Thus, the bounded convergence theorem applies. An example of such a model is shown in Figure 9. Whether this structure can fit observed cancer incidence data or not will be studied in Chapter 5.

Although the choice of a continuous distribution for γ seems intuitively natural, one can reasonably model γ by a discrete law. The multistage model is a complicated model, and for continuous distribution functions $F(\gamma)$ we get a mathematically untractable situation. Therefore, it is also interesting to assume a list $0 < \gamma_1 < \gamma_2 < \cdots < \gamma_k$ of possible values and work with the finite mixture model

$$S(t) = \sum_{i=1}^{k} \pi_i S(t|\gamma_i)\,.$$

We will see in Chapter 5 that such models yield good fits in practical applications. From the biological point of view, heterogeneity in γ has tangible consequences only if the differences are large enough. So it is crucial to know whether there are several groups with clearly changing γ values, and a discrete model seems appropriate for this purpose.

Finally, note that similar considerations could be made using the rate of malignant transformation, μ, instead of the pure growth rate γ.

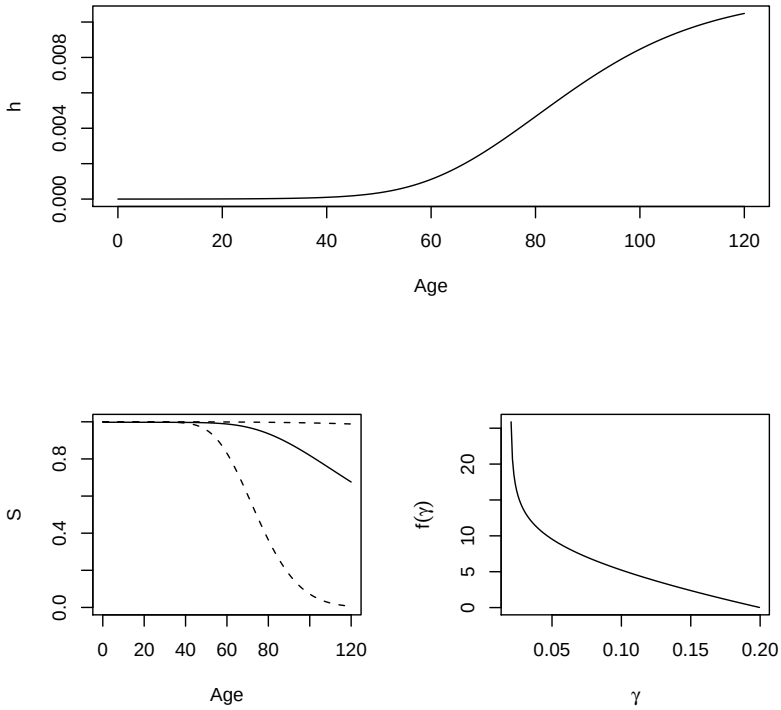

Figure 9: *Population hazard function $h(t)$ for the continuous multistage mixture model where γ is beta distributed with support $[\gamma_l, \gamma_u] = [0.02, 0.2]$ and shape parameters $a = 0.8$ and $b = 2$. The lower left panel shows the population survivor function $S(t)$ (solid line) along with $S(t|\gamma_l)$ and $S(t|\gamma_u)$ (dashed lines). The lower right panel gives the density $f(\gamma)$.*

CHAPTER 4

Identifiability

Since the first appearance of stochastic carcinogenesis models in the 1950s, it has taken quite a long time for the scientific community to understand the identifiability issue. Heidenreich (1996) showed that for the classical TSCE model the survivor function (and therefore also the hazard function) depends only on three parameters. This means the biological parameters are non-identifiable from time to tumor data alone. In Hanin and Yakovlev (1996) and Heidenreich et al. (1997) identifiable combinations of the biological parameters are given. Sherman and Portier (1997) discuss ways to overcome non-identifiability. They argue that either the parameter space must be reduced through fixing of some parameters or additional constraints, or extra types of data must be available. Hanin and Boucher (1999) and Hanin (2002) discuss a modification of the modeling assumptions and identifiability of this new model.

In this chapter we will give a series of identifiability results. First we will consider the multistage model as given in Section 2.3. We will

show that the results known from the TSCE model carry over to this slightly more general case. Then we will discuss identifiability of the mixture structure for selected finite mixture models.

4.1 The Multistage Model

Recall the definition of identifiability. Identifiability of a parametric model $\{\mathrm{P}_\theta\}_{\theta\in\Theta}$ means that $\mathrm{P}_{\theta_1} \neq \mathrm{P}_{\theta_2}$ whenever $\theta_1 \neq \theta_2$. This is a necessary condition for any sound statistical inference. The definition says that we must make sure that two different parameter values cannot yield the same probability measure.

The Number of Mutations for Initiation

Let us first focus on the number of mutations necessary for a cell to be initiated: is n identifiable in the multistage model? Could a change in n be compensated by some adjustment of the biological parameters ν, β, δ, μ, leading to the same survivor function? The following proposition shows that the behavior of $S(t|n)$ at the origin is enough to determine n.

Let $S(t|n,\psi)$ be the survivor function of the multistage model given in (2), where $\psi = (N_0, \nu, \beta, \delta, \mu)$ denotes the vector of biological parameters.

Proposition 2. *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: Let us define the integral

$$I(t;n) := \int_0^t \lambda_I(t-x;n) F_P(x)\,\mathrm{d}x\,.$$

Then, $S(t|n,\psi) = \exp\{-I(t;n)\}$. We omit ψ, since the proof will depend only on n, but not on the other biological parameters. We calculate the first $(n+1)$ derivatives of $I(t;n)$ with respect to t. For

any $n \geq 1$,

$$\begin{aligned}
I'(t;n) &= \int_0^t n(n-1)\nu^n N_0 (t-x)^{n-2} F_P(x)\,\mathrm{d}x\,, \\
I^{(2)}(t;n) &= \int_0^t n(n-1)(n-2)\nu^n N_0 (t-x)^{n-2} F_P(x)\,\mathrm{d}x\,, \\
&\vdots \\
I^{(n-1)}(t;n) &= \int_0^t n!\nu^n N_0 F_P(x)\,\mathrm{d}x\,, \\
I^{(n)}(t;n) &= n!\nu^n N_0 F_P(t)\,, \\
I^{(n+1)}(t;n) &= n!\nu^n N_0 F_P'(t)\,.
\end{aligned}$$

This shows that

$$I^{(k)}(0;n) = 0\,, \text{ for } k = 0, 1, 2, \ldots, n\,,$$

and $I^{(n+1)}(0;n) \neq 0$ if $f_P(0) = \frac{d}{dx}F_P(0) \neq 0$. It follows from equation (4) that $f_P(0) = \mu > 0$.

We can now express the derivatives of the survivor function $S(t|n)$ using the derivatives of $I(t;n)$. In order to do this, let $\mathcal{P}(x_1, \ldots, x_m)$ denote a real polynomial in the variables $x_1, \ldots, x_m$ (of arbitrary degree) without constant term. Using this notation, straightforward calculation shows that

$$S^{(k)}(t|n) = S(t|n) \cdot \mathcal{P}\left(I(t;n), I'(t;n), \ldots, I^{(k)}(t;n)\right)\,,$$

for $1 \leq k \leq n-1$, and

$$\begin{aligned}
S^{(n)}(t|n) &= -I^{(n)}(t;n)S(t|n) \\
&\quad + S(t|n) \cdot \mathcal{P}\left(I(t;n), I'(t;n), \ldots, I^{(n-1)}(t;n)\right)\,, \\
S^{(n+1)}(t|n) &= -I^{(n+1)}(t;n)S(t|n) \\
&\quad + S(t|n) \cdot \mathcal{P}\left(I(t;n), I'(t;n), \ldots, I^{(n)}(t;n)\right)\,.
\end{aligned}$$

Therefore,

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

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

In other words, to know how fast the survivor function $S(t|n, \psi)$ changes from its initial value $S(0|n) = 1$ as t increases is enough to determine n. However, this result is of theoretical interest and cannot be used in practice to estimate n.

The Biological Parameters

We will now consider the biological parameters ψ. As stated earlier, the case $n = 1$ corresponds to the TSCE model, and it has been shown in Hanin and Yakovlev (1996) that three functions of ψ are uniquely determined by $S(t|1, \psi)$. 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. The line of reasoning given by Hanin and Yakovlev can be generalized and the same result holds for any $n \geq 1$.

Lemma 1. *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: Note that by Proposition 2 both survivor functions, $S(t|n, \psi)$ and $S(t|\tilde{n}, \tilde{\psi})$, must be based on the same number of stages, that means $n = \tilde{n}$. So we must show that if

$$I(t; n, \psi) \equiv_t I(t; n, \tilde{\psi})\,, \tag{11}$$

then

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

First, we transform $I(t;n,\psi)$ via $(n-1)$ repeated integrations by parts into a n-fold integral,

$$\begin{aligned} I(t;n,\psi) &= n\nu^n N_0 \int_0^t (t-u_1)^{n-1} F_P(u_1)\,\mathrm{d}u_1 \\ &= n(n-1)\nu^n N_0 \int_0^t (t-u_1)^{n-2} \int_0^{u_1} F_P(u_2)\,\mathrm{d}u_2\,\mathrm{d}u_1 \\ &\vdots \\ &= n!\nu^n N_0 \int_0^t \int_0^{u_1} \dots \int_0^{u_{n-1}} F_P(u_n)\,\mathrm{d}u_n \dots \mathrm{d}u_1\,. \end{aligned}$$

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 (11) proves the result.□

Lemma 1 generalizes a result by Hanin and Yakovlev that is used to show non-identifiability of the two-stage model. The rest of their argument can directly be applied to our case. More precisely, this means that $S(t|n,\psi)$ uniquely defines the three parameter combinations

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

In the sequel, we will therefore fix the two parameters N_0 and δ to get an identifiable model. Note that for example N_0 =*number of stem cells in a given tissue* and $\delta = 0$ is an attractive model of this type. We will also focus on $\gamma = \beta - \delta$ rather than on β itself, since it is the growth advantage of initiated cells that is of real biological interest.

4.2 Identifiability of Mixture Models

We will now turn attention to the identifiability of the mixture structure itself. Let us review briefly the necessary mathematical theory, which

will be useful for the study of some mixture models of the multistage model in the next section.

First of all, recall the definition of identifiability of a mixture model. Let $S(t|\theta)$ be a survivor function depending on a parameter θ, and let $\mathcal{G}$ be a set of probability distribution functions for θ. Then, $\mathcal{G}$ induces the set of mixture models

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

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

$$\int S(t|\theta)\, \mathrm{d}G_1(\theta) \equiv_t \int S(t|\theta)\, \mathrm{d}G_2(\theta) \implies G_1 \equiv_\theta G_2\,, \qquad (12)$$

for all $G_1, G_2 \in \mathcal{G}$. The definition says that if we know the population survivor function

$$S(t) = \int S(t|\theta)\, \mathrm{d}G(\theta)\,,$$

then we know also the underlying mixing distribution $G(\theta)$ that has generated it, at least within a given set of candidate distributions, $\mathcal{G}$. Very often we will simply say that the mixture model $S(t)$ is identifiable with respect to $\mathcal{G}$.

Let us first focus on the case of finite mixture models. In this setting, the survivor functions are of the form

$$S(t) = \sum_{i=1}^{k} \pi_i S(t|\theta_i)\,,$$

with weights such that $0 < \pi_i < 1$, and $\sum_{i=1}^{k} \pi_i = 1$. The number of components $k \in \mathbb{N}$ must not necessarily be fixed, and $\theta_1, \ldots, \theta_k$ are supposed to belong to a given countable set $\{\vartheta_1, \vartheta_2, \vartheta_3, \ldots\}$. Let $S_1(t) = \sum_{i=1}^{k} \pi_i S(t|\theta_i)$ and $S_2(t) = \sum_{i=1}^{k'} \pi'_i S(t|\theta'_i)$ be two mixture models. The identifiability condition (12) now becomes

$$S_1(t) \equiv_t S_2(t) \implies \{(\pi_1, \theta_1), \ldots, (\pi_k, \theta_k)\} = \{(\pi'_1, \theta'_1), \ldots, (\pi'_{k'}, \theta'_{k'})\}\,.$$

This means the summands of the mixture are uniquely determined up to permutations. In order to eliminate this ambiguity, one must require a supplementary condition. For example one can require that $\vartheta_1 < \vartheta_2 < \vartheta_3 < \dots$.

In this work, we will use extensively the following result by Teicher (1963):

Let $\{S_i(t)\,;\ i = 1, 2, \dots\}$ *be a family of survivor functions such that*

1) $S_i(t) > 0\,,\ \forall\ t \geq 0\,,\ i = 1, 2, \dots\,;$

2) $\exists\ a \in \mathbb{R}^+ \cup \{\infty\}$ *such that* $\lim_{t\to a} \frac{S_{i+1}(t)}{S_i(t)} = 0\,,\ i = 1, 2, \dots$.

Then, the finite mixture model $S(t) = \sum_{i=1}^{k} \pi_i S_i(t)$ *is identifiable.*

In order to establish the result, one first notes that for any positive integer l , condition *2)* implies

$$\lim_{t\to a} \frac{S_{i+l}(t)}{S_i(t)} = \lim_{t\to a} \frac{S_{i+l}(t)}{S_{i+l-1}(t)} \cdot \frac{S_{i+l-1}(t)}{S_{i+l-2}(t)} \cdots \frac{S_{i+1}(t)}{S_i(t)} = 0\,.$$

Let now be given two identical finite mixtures

$$\sum_{i=1}^{k} \pi_i S_{n_i}(t) \equiv_t \sum_{j=1}^{k'} \pi'_j S_{m_j}(t)\,, \tag{13}$$

such that again

$$0 < \pi_i, \pi'_j < 1\,,\ \sum_{i=1}^{k} \pi_i = \sum_{j=1}^{k'} \pi'_j = 1\,. \tag{14}$$

Without loss of generality one can assume that

$$n_1 < n_2 < \dots < n_k \ \text{ and } \ m_1 < m_2 < \dots < m_{k'}\,.$$

If $n_1 \neq m_1$, again without loss of generality one can assume that $n_1 < m_1$. Then, by (13),

$$\pi_1 + \sum_{i=2}^{k} \pi_i \frac{S_{n_i}(t)}{S_{n_1}(t)} \equiv_t \sum_{j=1}^{k'} \pi'_j \frac{S_{m_j}(t)}{S_{n_1}(t)}\,,$$

and letting $t \to a$ one gets $\pi_1 = 0$, in contradiction to (14). Therefore $n_1 = m_1$, and

$$\pi_1 - \pi_1' + \sum_{i=2}^{k} \pi_i \frac{S_{n_i}(t)}{S_{n_1}(t)} \equiv_t \sum_{j=2}^{k'} \pi_j' \frac{S_{m_j}(t)}{S_{n_1}(t)} .$$

Letting again $t \to a$ yields $\pi_1 - \pi_1' = 0$ and equation (13) reduces to

$$\sum_{i=2}^{k} \pi_i S_{n_i}(t) \equiv_t \sum_{j=2}^{k'} \pi_j' S_{m_j}(t) .$$

Repeating this argument a finite number of times it follows that

$$n_i = m_i , \ \pi_i = \pi_i' , \text{ for } i = 1, \ldots, \min\{k, k'\} .$$

If $k \neq k'$, e.g. $k > k'$, then again by (13),

$$\sum_{i=k'+1}^{k} \underbrace{\pi_i}_{>0} \underbrace{S_{n_i}(t)}_{>0} \equiv_t 0 ,$$

which is a contradiction. Thus, $k = k'$. Teichers theorem is very useful in practice, since it links identifiability to a condition that is easy to check in many applications. In the sequel, we will refer to this result as *Teichers condition*.

As we might expect, the continuous case is much more difficult to treat. Results exist for special classes of distributions (like additively closed families, the translation or the scale parameter family, Gaussian distributions, exponential distributions), or for special models. Identifiability conditions from the general theory of integral equations are given by Tallis and Chesson (1982). However, their results are hard to apply in practice.

Finally, let us make a remark about identifiability in the classical frailty modeling framework. Most of the frailty literature treats the model

$$\mathrm{h}_{\mathrm{ind}}(t) = Z \cdot \mathrm{h}_0(t) ,$$

where the frailty Z is a non-negative random variable. This leads to a population hazard

$$\mathrm{h}_{\mathrm{pop}}(t) = \mathrm{h}_0(t) \cdot \mathrm{E}(Z|T > t)\,.$$

The model is non-identifiable unless assumptions on the amount of variability among individuals are made. For example, one often puts $\mathrm{E}(Z) = 1$ (see Hougaard (1995)).

4.3 Finite Mixtures of the Multistage Model

Initiation

In Chapter 3 we proposed a finite mixture model based on the idea of germ line mutations in cancer causing genes. The population survivor function was

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

The next proposition shows that such a mixture is identifiable.

Proposition 3. *The family of finite mixture models induced by*

$$\{S(t|n,\psi)\,;\ n = 1, 2, \dots\}$$

is identifiable.

Proof: We show that Teichers condition is satisfied. Part *1)* holds since $S(t|n,\psi)$ is the composition with an exponential function. In order to check part *2)*, let us study the ratio

$$\frac{S(t|n+1,\psi)}{S(t|n,\psi)} = \exp\{-\int_0^t [\lambda_I(x;n+1) - \lambda_I(x;n)] F_P(t-x)\,\mathrm{d}x\}\,.$$

The initiation incidence rates can be written recursively

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

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

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

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

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

□

Promotion

In Chapter 3 we also studied a finite mixture model that was motivated by variation in γ. A similar result as above can be derived for this case. Let there be a discrete set of γ-values, $0 < \gamma_1 < \gamma_2 < \dots$. Note that we consider $\delta = \gamma_i - \beta_i$ as fixed, that is, we assume in fact that there is an analogous sequence of β_i. We will write ψ for the parameter vector $(n, N_0, \nu, \delta, \mu)$.

Proposition 4. *The family of finite mixture models induced by*

$$\{S(t|\gamma_i,\psi)\,;\ i = 1,2,\dots\}$$

is identifiable.

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

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

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. The assumption $\gamma_{i+1} > \gamma_i$ implies that $a(\gamma_{i+1}, \delta) > a(\gamma_i, \delta)$, and as a consequence

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

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

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

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

$$\frac{\partial}{\partial\beta}F_P(x|\beta) = \frac{\mu e^{-(\beta+\mu)x}[(\beta+\mu)x + e^{-(\beta+\mu)x} - 1]}{[\mu + \beta e^{-(\beta+\mu)x}]^2}\,.$$

Therefore, we have

1) $\frac{\partial}{\partial\beta}F_P(0|\tilde\beta) = 0$;

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

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

This shows that

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

which completes the proof. □

The same ideas 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.

Counterintuitive Properties

The proof of proposition 3 reveals the somehow counterintuitive fact that $S(t|n+1,\psi) > S(t|n,\psi)$ does not hold for all t. Logically, individuals needing $n+1$ events for initiation must have a higher probability to survive up to age t than those needing only n such events. And of course, the model does have this property for all reasonable lifetimes t. However, far in the tails the two survivor functions cross. Let us illustrate this for the simpler multi-hit model $S_{\text{MH}}(t)$ given in Section 2.2. Recall the survivor function

$$S_{\text{MH}}(t|n) = \exp\{-N_0 \nu^n t^n\}\,.$$

We do have as expected $S_{\text{MH}}(t|n) \leq S_{\text{MH}}(t|n+1)$ for $t \in [0, \nu^{-1}]$. But if $t > \nu^{-1}$, then the order is reversed. Note that ν is interpreted as a mutation rate per year, i.e. ν is a very small number (of about the order 10^{-5}).

The inconsistent behaviour in the tails is a consequence of the approximation used to derive the survivor function $S_{\text{MH}}(t)$. To illustrate this, let us consider the exact solution for a multi-hit process with n hits. Let X_i denote the waiting time for a single cell to pass from compartment $i-1$ to compartment i. We assume the sequence $X_1, \ldots, X_n$ be independent and identically exponentially distributed $\mathcal{E}(\nu)$. Then, the waiting time for a single cell to be transformed into a tumor cell,

$$T_n = X_1 + \cdots + X_n\,,$$

is gamma distributed with shape parameter n and scale parameter $1/\nu$. Let there be N_0 cells at risk in the organism. We assume N_0 be constant over time and suppose the different cells at risk to behave independently of one another. Then, we can attribute such a waiting time to every of the N_0 cells and get the sequence of independent and identically distributed random variables $T_n^1, \ldots, T_n^{N_0}$. With this notation, the survivor function of the organism for tumor onset is

$$\tilde{S}_{\text{MH}}(t|n) = \mathrm{P}(\min_j T_n^j > t) = \mathrm{P}(T_n > t)^{N_0} .$$

This notation shows that

$$\tilde{S}_{\text{MH}}(t|n) < \tilde{S}_{\text{MH}}(t|n+1) , \ \forall\, t > 0 .$$

The survival probability is indeed higher if more hits are needed. The inconsistency disappears.

If we want to study the fraction

$$\frac{\tilde{S}_{\text{MH}}(t|n)}{\tilde{S}_{\text{MH}}(t|n+1)} = \frac{\mathrm{P}(T_n > t)^{N_0}}{\mathrm{P}(T_{n+1} > t)^{N_0}} ,$$

we see that it is enough to consider a single cell, that is the fraction with N_0 replaced by 1. Since T_n is gamma distributed, we get

$$\frac{\mathrm{P}(T_n > t)}{\mathrm{P}(T_{n+1} > t)} = \frac{n\Gamma(n, \nu t)}{\Gamma(n+1, \nu t)} ,$$

where

$$\Gamma(n, t) = \int_t^\infty e^{-x} x^{n-1} \, \mathrm{d}x$$

is the upper incomplete gamma function. Since n is an integer, we have the representation

$$\Gamma(n, t) = (n-1)!\, e^{-t} \sum_{k=0}^{n-1} \frac{t^k}{k!} .$$

Therefore, we get

$$\frac{\mathrm{P}(T_n > t)}{\mathrm{P}(T_{n+1} > t)} = \frac{1 + \nu t + \cdots + (\nu t)^{n-1}/(n-1)!}{1 + \nu t + \cdots + (\nu t)^n/n!} \leq 1 , \ \forall\, t \geq 0 ,$$

with equality if and only if $t = 0$. The above expression implies also that

$$\frac{\tilde{S}_{\text{MH}}(t|n)}{\tilde{S}_{\text{MH}}(t|n+1)} \xrightarrow{t \to \infty} 0\,, \ \forall\, n\,.$$

Applying Teichers condition, we have shown the following

Proposition 5. *The family of finite mixture models induced by*

$$\{\tilde{S}_{MH}(t|n)\,;\ n = 1, 2, \dots\}$$

is identifiable.

CHAPTER 5

Application to Data

5.1 Data Description

We will now apply our mixture models to human cancer incidence data. Such data typically shows a sharp increase in the number of cases from about the age of forty, reaches a peak around the age of eighty, and decreases for the very old. This effect can be explained by the dying out of the high-risk group of the population. In order to study this phenomenon, we need incidence data from a population with individuals that have lived under comparable conditions. Therefore, we need birth cohort rather than cross-sectional data. We will investigate lung cancer and colon cancer data for European Americans born in the 1880s, 1890s, 1900s, and 1920s from Herrero-Jimenez et al. (2000) and Morgenthaler et al. (2004).

The data actually consists of population counts from 1930 to 1991, of mortality tables from lung cancer for the same period, and of mortality tables from colon cancer from 1958 to 1991. In all cases the counts

were grouped into five year age intervals: 0-4 years, 5-9 years, 10-14 years and so on. For simplicity we assume these binned counts to be equally distributed within their age groups to obtain the cohort data.

In order to apply a carcinogenesis model, we must adjust the data for the confusion of mortality and incidence. The data at hand gives only site specific mortality, while the stochastic models discussed in Chapters 2 to 4 describe incidence. Mortality does reflect incidence, but will underestimate the number of cancer cases. This is a minor issue for highly fatal tumors, where mortality is approximately equal to incidence. But in many other cases, there is relatively efficient treatment available, and some kind of correction must be applied. Let $\tilde{o}_i$ be the number of deaths during the interval $[t_i, t_{i+1})$. Then, we will simply multiply $\tilde{o}_i$ by a constant to obtain the observed incidence, $o_i = c \cdot \tilde{o}_i$. For lung cancer, patients still have very poor prospects. In Stewart and Kleihues (2003) the worldwide numbers of lung cancer cases and deaths per year are given. These counts, reproduced in Table 1, can be used to obtain a reasonable value for c.

		Males	Females
Lung Cancer	Incidence	901	337
	Mortality	810	292

Table 1: *Worldwide numbers of lung cancer cases and deaths in thousands per year given by the WHO (Stewart and Kleihues (2003)).*

This leads to the sex specific correction factors $c_m = 1.11$, $c_f = 1.15$. These numbers are consistent with survivor probabilities given in US registries. We obtain the adjusted observed incidence data shown in Table 2. The corresponding hazard curves are given in Figure 10. Note that in Table 2 the population at risk for the 1890s, the 1900s, and the 1920s group is not strictly decreasing. This is due to immigration. The data in Table 2 is, however, the best approximation to human cohort data we could get. Whenever the increase in the at risk sets is

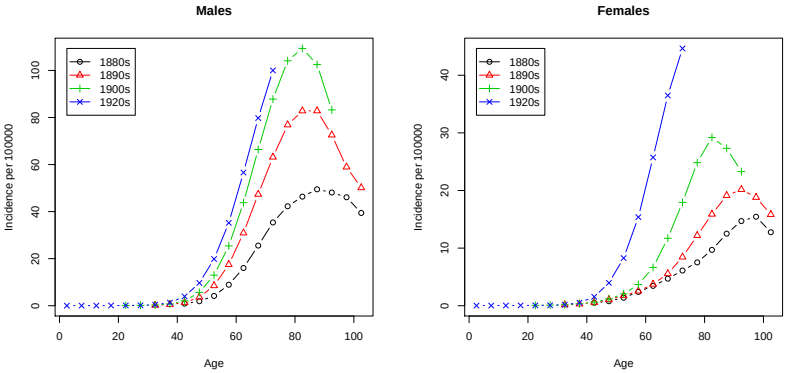

Figure 10: *Observed lung cancer incidence corresponding to the data given in Table 2. The raw hazard estimates are calculated* $\hat{\lambda}_i = \frac{o_i}{r_i(t_{i+1}-t_i)}$ *, where* r_i *counts the population at risk and* o_i *the number of cases over the interval* $[t_i, t_{i+1})$ *.*

incompatible with a method, we will restrict the analysis to the 1880s birth cohort.

The case of colon cancer is more difficult to treat. The cure rate is much higher, and the treatments evolved considerably over the time span we consider. In Herrero-Jimenez et al. (2000) this problem is discussed. These authors estimate the survival rates for colon cancer by age and year of diagnosis. Their results are given in Table 3. Let $\tilde{o}(h,t)$ denote the number of deaths from birth cohort h at age t. We can then estimate the actual number of colon cancer cases as

$$o(h,t) = \tilde{o}(h,t)/(1 - S(h,t)) \,,$$

where $S(h,t)$ is the relative colon cancer survival rate for age cohort h at age t (read out of Table 3). We get the adjusted colon cancer incidence data shown in Table 4 and Figure 11.

Let us finally mention that often only cross-sectional data is available. In this case, one can take the multistage model as age-specific

Males								
	1880s		1890s		1900s		1920s	
Age	At risk	Cases	At risk	Cases	At risk	Cases	At risk	Cases
[0, 5)							49475170	68
[5, 10)							51880815	90
[10, 15)							52755426	128
[15, 20)							53422234	256
[20, 25)					46424286	274	52475231	316
[25, 30)					44910182	378	54795743	428
[30, 35)			40683682	407	44667053	637	55221754	1206
[35, 40)			41873241	1083	45202500	1687	55294801	3673
[40, 45)	37177005	1677	39700013	2699	45707353	4739	55008545	10906
[45, 50)	35222529	3362	39592494	7163	44518021	12610	54123447	26223
[50, 55)	33219575	6927	37969964	16161	42964495	27904	52411957	51993
[55, 60)	30583428	13659	36270150	31740	40410105	51346	49739071	87606
[60, 65)	28701833	23011	32952183	50956	36610956	80189	46125810	130597
[65, 70)	25400588	32468	28347996	67129	31535474	104682	41230912	164468
[70, 75)	20249992	35839	22607010	71395	25621407	112494	33816110	169079
[75, 80)	14526850	30718	16288426	62592	19258373	100255		
[80, 85)	8851802	20525	10314523	42708	12539867	68564		
[85, 90)	4247329	10508	5365053	22220	6136070	31448		
[90, 95)	1527841	3676	2080776	7549	2284233	9505		
[95, 100)	373936	862	503896	1485				
[100, 105)	54275	107	66610	167				

Females								
Age	1880s		1890s		1900s		1920s	
	At risk	Cases	At risk	Cases	At risk	Cases	At risk	Cases
[0, 5)							47616208	54
[5, 10)							50281978	64
[10, 15)							51160245	89
[15, 20)							52612257	144
[20, 25)					47573246	170	53078675	171
[25, 30)					45740263	237	55292676	207
[30, 35)			40618853	328	44994177	389	55700064	474
[35, 40)			40880241	591	45260577	776	55881276	1451
[40, 45)	34845993	877	38463333	1120	45620766	1415	55792798	4342
[45, 50)	32927685	1291	38474741	2031	44891751	2669	55338368	10979
[50, 55)	30962992	2146	37936954	3312	44213190	4522	54388589	22500
[55, 60)	29497353	3539	37563639	4781	42985653	7912	52875595	40645
[60, 65)	28978869	4998	35826007	6688	41114123	13679	51045641	65700
[65, 70)	27200715	6387	33221776	9238	38408131	22477	48937816	89287
[70, 75)	23757908	7258	29466608	12445	34742896	31152	43665837	97489
[75, 80)	19234172	7235	24495607	14921	30012287	37272		
[80, 85)	13755533	6667	18597769	14772	23620919	34490		
[85, 90)	8080774	5054	12090697	11558	14998028	20476		
[90, 95)	3704766	2722	6116973	6168	7288050	8483		
[95, 100)	1206871	932	1903266	1789				
[100, 105)	242950	155	329783	261				

Table 2: *Lung cancer incidence data from European Americans.*

Year of diagnosis		Ages 0-44	45-54	55-64	65-74	75+	100+
1990s	m	0.58	0.62	0.65	0.66	0.59	0.03
	f	0.59	0.65	0.62	0.64	0.60	0.04
1980s	m	0.49	0.59	0.59	0.60	0.57	0.03
	f	0.58	0.56	0.56	0.57	0.55	0.04
1970s	m	0.47	0.48	0.48	0.48	0.44	0.03
	f	0.58	0.50	0.50	0.48	0.46	0.04
1960s	m	0.50	0.45	0.45	0.44	0.37	0.03
	f	0.50	0.48	0.48	0.47	0.42	0.04
1950s	m	0.42	0.46	0.40	0.38	0.32	0.03
	f	0.46	0.46	0.46	0.42	0.38	0.04
1940s	m	0.27	0.33	0.29	0.21	0.17	0.03
	f	0.34	0.34	0.35	0.28	0.24	0.04
1930s	m	0.30	0.27	0.20	0.09	0.00	0.03
	f	0.25	0.17	0.18	0.11	0.07	0.04

Table 3: *Relative survival rates for colon cancer by age and year of diagnosis from Herrero-Jimenez et al. (2000).*

baseline hazard $h(t)$. An adjustment for the birth year i and for the year of diagnosis j must then be applied. In Luebeck and Moolgavkar (2002), the model $h_{ij}(t) = b_i c_j h(t)$ is considered. In particular, these authors investigate the evolution of b_i and c_j over time.

Males								
	1880s		1890s		1900s		1920s	
Age	At risk	Cases	At risk	Cases	At risk	Cases	At risk	Cases
[25, 30)							54795743	438
[30, 35)							55221754	1034
[35, 40)							55294801	2056
[40, 45)							55008545	4494
[45, 50)					44518021	5731	54123447	8195
[50, 55)					42964495	12451	52411957	16760
[55, 60)			36270150	14761	40410105	18184	49739071	29530
[60, 65)			32952183	24894	36610956	31072	46125810	57887
[65, 70)	25400588	24000	28347996	32382	31535474	43153	41230912	80411
[70, 75)	20249992	35138	22607010	43779	25621407	58148	33816110	105098
[75, 80)	14526850	29669	16288426	40205	19258373	56166		
[80, 85)	8851802	24846	10314523	38042	12539867	60703		
[85, 90)	4247329	14600	5365053	24143	6136070	36335		
[90, 95)	1527841	6647	2080776	13662	2284233	16024		
[95, 100)	373936	1657	503896	3324				
[100, 105)	54275	137	66610	162				

Females								
	1880s		1890s		1900s		1920s	
Age	At risk	Cases	At risk	Cases	At risk	Cases	At risk	Cases
[25, 30)							55292676	439
[30, 35)							55700064	1281
[35, 40)							55881276	2885
[40, 45)							55792798	5667
[45, 50)					44891751	8061	55338368	9830
[50, 55)					44213190	15909	54388589	17151
[55, 60)			37563639	17446	42985653	22962	52875595	26882
[60, 65)			35826007	29722	41114123	32946	51045641	42324
[65, 70)	27200715	27986	33221776	36678	38408131	45614	48937816	58616
[70, 75)	23757908	39534	29466608	49928	34742896	59122	43665837	84852
[75, 80)	19234172	38550	24495607	52121	30012287	65247		
[80, 85)	13755533	36850	18597769	55591	23620919	76709		
[85, 90)	8080774	26956	12090697	43701	14998028	60756		
[90, 95)	3704766	13712	6116973	28950	7288050	37817		
[95, 100)	1206871	4371	1903266	9335				
[100, 105)	242950	406	329783	559				

Table 4: *Colon cancer incidence data from European Americans.*

5.2 Finite Mixtures and Lung Cancer Data

We will now apply our mixture models to the lung cancer data shown above. In a first step, we consider the easiest situation of a two component mixture. Most of the time, we use the γ-frailty model with survivor function

$$S(t) = \pi_l S(t|\gamma_l) + (1 - \pi_l) S(t|\gamma_u)\,, \tag{15}$$

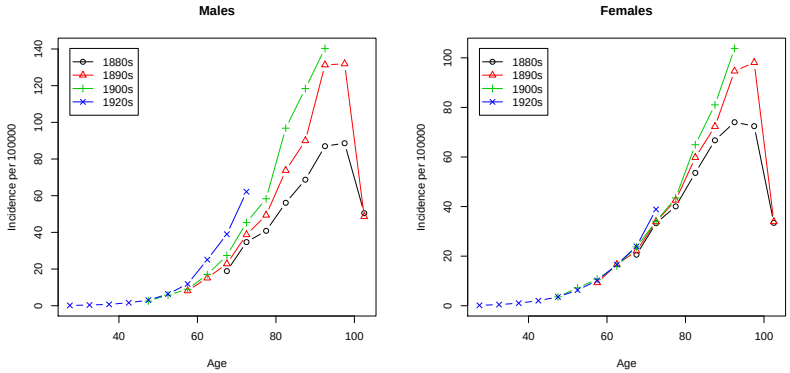

Figure 11: *Observed lung cancer incidence corresponding to the data given in Table 4.*

where $0 < \gamma_l < \gamma_u$, $0 < \pi_l < 1$, and $S(t|\gamma)$ is the survivor function (2) of the multistage carcinogenesis model given in Section 2.3. Model (15) will serve to illustrate the main issues concerning parameter estimation, and its choice is somewhat arbitrary. Other frailty models will be considered later.

Recall the complete list of parameters of model (15): N_0, the number of cells at risk; n, the number of mutations necessary to initiate a cell; ν and μ, the rates of initiating mutations and malignant transformation; γ_l and γ_u, the net growth rates of initiated cells; δ, the death rate of initiated cells; and π_l, the proportion of the population with low net cell growth rate. In order to guarantee identifiability, we will fix N_0 and δ. Since the parameters are biologically meaningful, we can find reasonable values for them in the literature. But we must keep in mind that all estimates will be conditional given N_0 and δ. The former of these two acts as a scale parameter. Changes in N_0 are compensated by ν, and qualitatively no different effects arise for altering N_0 values. The role of δ, however, is more difficult to assess, and it is reasonable to perform some sensitivity analysis.

In order to get stable estimates, we will usually fix some more parameters. For model (15) for example, we will estimate $(\pi_l, \nu, \gamma_u, \mu)$ for several choices of previously specified n and γ_l. Numerical optimization will be done using the simplex method described in Nelder and Mead (1965). The procedure is implemented in the `R` function `optim`. This method does not require derivatives, which is an important feature in our case, since $S(t)$ is a rather complicated function in terms of the biological parameters. The simplex method is relatively slow, but robust.

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. More precisely, numerical optimization will always be performed using the following reparametrization: $\tilde{\pi}_l = \text{logit}(\pi_l)$, $\tilde{\gamma} = \log(\gamma_u - \gamma_l)$, $\tilde{\nu} = \log_{10}(\nu)$, and $\tilde{\mu} = \log_{10}(\mu)$. Recall the definition of the logit-function,

$$\begin{array}{rccc} \text{logit:} & (0,1) & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \ln\left(\frac{x}{1-x}\right), \end{array}$$

which we used to transform the probability π_l.

5.2.1 Maximum Likelihood

The first procedure we consider is maximum likelihood estimation. In order to derive the likelihood function for our case, it is useful to represent the cohort data on a time axis as in Figure 12.
We see that the cohort is made up by r_0 individuals. But instead of the lifetimes $y_1, \ldots, y_{r_0}$, we only know how many persons failed in $[t_i, t_{i+1})$ due to cancer, namely o_i, and how many failed due to competing causes, namely

$$c_i = r_i - r_{i+1} - o_i .$$

We will treat these cases as right censorings. Thus, we get two types of likelihood contributions from the totally $o_i + c_i$ failures during $[t_i, t_{i+1})$:

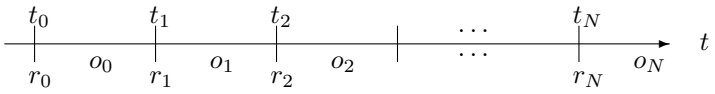

Figure 12: *Our data structure. At fixed time points* $0 \leq t_0 < t_1 < \cdots < t_N$ *we know the population at risk* $r_0\,,\ r_1\,,\ldots,\ r_N\,,$ *while* $o_0\,,\ o_1\,,\ldots,\ o_N$ *denote the failures due to cancer in the intervals between two consecutive time points.*

the o_i individuals having a cancer give each a factor $(S(t_i) - S(t_{i+1}))\,,$ while the c_i censored individuals contribute $S(t_i)\,.$ This means, we obtain the likelihood

$$L(\pi_l, \nu, \gamma_u, \mu | N_0, \delta, n, \gamma_l) = \prod_{i=0}^{N} \big(S(t_i) - S(t_{i+1})\big)^{o_i} S(t_i)^{c_i}\,.$$

This means we assume independent and noninformative censoring. We must adopt some convention for $t_{N+1}\,.$ We can either set $t_{N+1} = \infty\,,$ in which case $S(t_{N+1}) = 0\,,$ or we can set t_{N+1} equal to some artificial finite endpoint. Note that in our applications both possibilities yield virtually the same results. But in some theoretical contexts it may be a necessary assumption that $t_{N+1} < \infty\,.$ Finally, we always set $r_{N+1} = 0\,.$

In order to estimate the biological parameters, we will use the previously introduced parametrization. That means we maximize the log-likelihood

$$l(\tilde{\pi}_l, \tilde{\nu},\, \mathrm{d}\tilde{\gamma}, \tilde{\mu}) = \sum_{i=0}^{N} \big\{ o_i \log\big(S(t_i) - S(t_{i+1})\big) + c_i \log S(t_i) \big\}\,.$$

Let us consider the lung cancer incidence data from the males 1880s cohort. Although our numerical optimizer converges, we observe a strange behaviour of the MLE. Figure 13 shows the data along with the models corresponding to the MLE, the LSE, and the starting value of the numerical optimization. As we can see, the MLE fails completely to catch the behaviour 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 behaviour 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 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.

Another point concerns the way we look at our data. In carcinogenesis, it is natural to work with hazard curves. On the population level, site specific cancer is a rare disease and survivor functions are extremely flat over the whole lifespan. Therefore, plots of survivor functions would be very hard to interpret; the corresponding hazard curves on the other hand show interesting patterns. The LSE is optimal with respect to fitting the observed hazard, while the likelihood we have introduced above uses the survivor function as main ingredient. Because $h(t) = -\,\mathrm{d}\log S(t)/\,\mathrm{d}t$, two hazard curves can be markedly different, although the corresponding survivor functions are rather close.

The above mentioned points are just two aspects of the problem. We investigate other issues in more detail below.

The Likelihood Surface

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

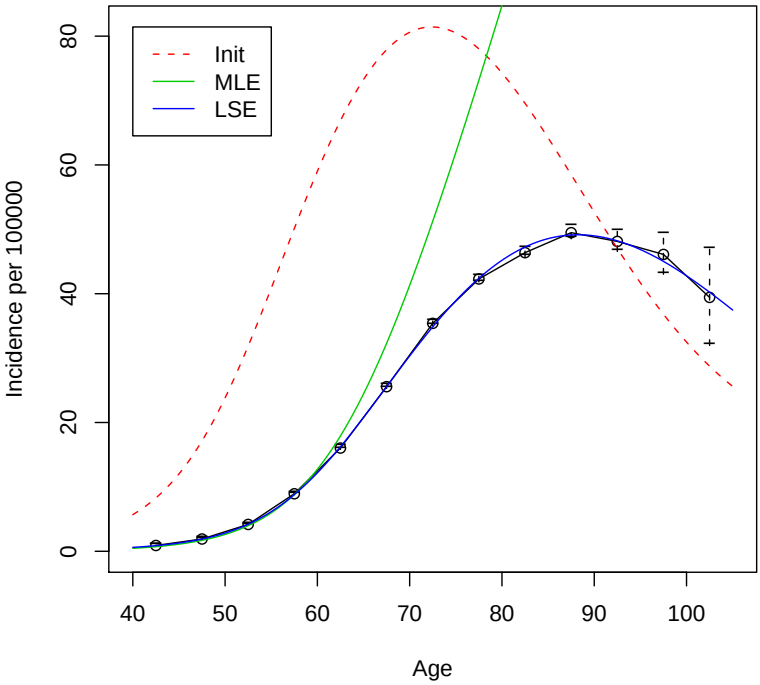

Figure 13: *Lung cancer data from the males 1880s cohort along with estimated models of the form (15). We fixed* $N_0 = 10^{10}$, $\delta = 0$, $n = 2$ $\gamma_l = 10^{-4}$ *, and used the initial value* $\tilde{\pi}_l = \text{logit}(0.97)$, $d\tilde{\gamma} = \log(0.2)$, $\tilde{\nu} = -6.5$ *and* $\tilde{\mu} = -5$.

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 14 gives the contour plot of the log-likelihood for a reduced parameter space. That is we take model (15), but fix all parameters except $\psi = (\tilde{\pi}_l, \tilde{\mu})$. We see that the log-likelihood surface has a ridge. While $\tilde{\mu}$ can be estimated relatively precisely, there is much more uncertainty about $\tilde{\pi}_l$. The log-likelihood values of the estimates in Figure 14 are

$$l(\hat{\psi}_{\text{ML}}) \simeq -1.338 \cdot 10^6\,,\ l(\hat{\psi}_{\text{LS}}) \simeq -1.355 \cdot 10^6\,.$$

This shows that on a relative scale, the two are extremely close,

$$l(\hat{\psi}_{\text{ML}})/l(\hat{\psi}_{\text{LS}}) \simeq 0.99\,.$$

But if we do a likelihood ratio test for $\hat{\psi}_{\text{LS}}$ against $\hat{\psi}_{\text{ML}}$, we get

$$2(l(\hat{\psi}_{\text{ML}}) - l(\hat{\psi}_{\text{LS}})) \simeq 32600 \gg q\chi^2_2(0.999) \approx 13.8\,,$$

which means that $\hat{\psi}_{\text{ML}}$ is significantly different from $\hat{\psi}_{\text{LS}}$. A 95% confidence region determined by a likelihood-ratio test is given in Figure 15. This set is actually extremely small; expressed in the natural parametrization (π_l, μ) it is contained in the rectangle $[0.944, 0.947] \times [8 \cdot 10^{-7}, 8.5 \cdot 10^{-7}]$.

Heavy Censoring

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

$$\tilde{r}_i^k = 10^6 - i \cdot k10^4\,, \quad \text{and} \quad \frac{\tilde{o}_i^k}{\tilde{r}_i^k} = \frac{o_i}{r_i}\,. \tag{16}$$

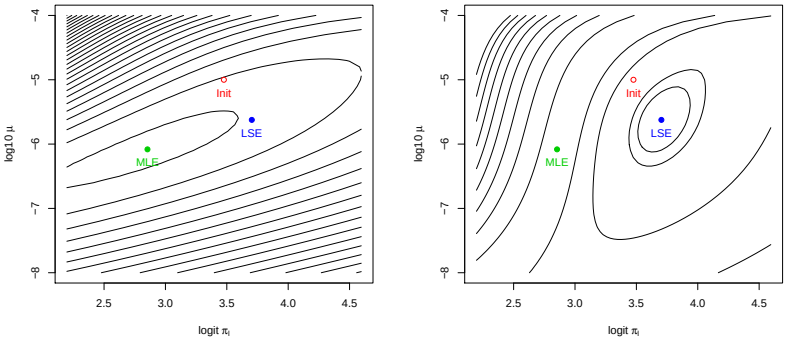

Figure 14: *Contour plot of the log-likelihood surface (left panel) and of the residual sums of squares surface. The parameter space is reduced such that only $\tilde{\pi}_l$ and $\tilde{\mu}$ are estimated. The other parameters are fixed as in Figure 13. Additionally, we set $\gamma_u = 0.192$ and $\nu = 2.43 \cdot 10^{-7}$, which are their respective least squares estimates from the unrestricted case.*

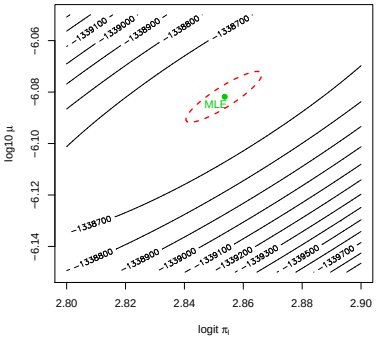

Figure 15: *Contour plot of the log-likelihood surface. The red dashed line gives the border of a 95% confidence region determined by a likelihood-ratio test.*

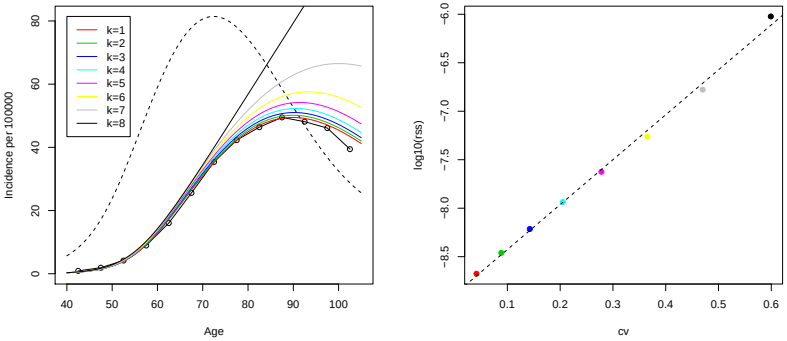

Figure 16: *Hazard curves corresponding to the MLEs for the data sets constructed according to (16). The right panel shows the residual sums of squares between the models and the raw hazard estimates as a function of the coefficient of variation of the at risk sets. Note that for the real data set we have* $cv(r_0, \ldots, r_{12}) \approx 0.77$.

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 (15) by maximum likelihood as before (we consider again the four parameters $\tilde{\pi}_l$ $\tilde{\nu}$ $\mathrm{d}\tilde{\gamma}$ $\tilde{\mu}$ as unknown). Figure 16 gives the estimated models for $k = 1, \ldots, 8$. The MLE does indeed behave better for small k than for large k. We also calculated the residual sums of squares (RSS) for these models, which seems 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. We have also seen that the LSE, on the contrary, 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 13 once again, we realize that there is a model that fits all the data points very accurately. This model will be extremely 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.

5.2.2 Analytic Graduation

The least squares estimate in the previous illustrations was obtained by analytic graduation described in Section 1.3.4. This method works very well in our application. Therefore, we will use this technique to analyze the different data sets introduced at the beginning of this chapter. In a first part, we will continue to use the model with γ-frailty. But later on, we will consider heterogeneity in other biological parameters as well.

Random Growth Advantage

Let us consider model (15) more closely. We mentioned earlier that we fixed the net growth rate of the low risk group, γ_l, in order to get stable estimates. The model corresponding to $\hat{\psi}_{\text{LS}}$ fitted the observed hazard for the lung cancer incidence data (males 1880s cohort) extremely accurately. But we must keep in mind that $\hat{\psi}_{\text{LS}}$ is conditional on γ_l. To assess the sensitivity of $\hat{\psi}_{\text{LS}}$ to γ_l, we repeated the estimation for several γ_l values. The result is given in Table 5. It turns out that we get good fits only when γ_l is small enough. This means that the large majority of the population must be at very low cancer risk. Consequently, the observed peak in the hazard curve is due to the dying out of the group of proportion $1-\pi_l$ with net growth rate γ_u. Table 5 also gives the results for several δ, but this seems to have a minor impact on parameter estimates.

We will now apply model (15) to the lung cancer incidence data from all cohorts and both sexes. The model is flexible enough to fit all the

δ	γ_l	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}$	$\hat{\mu}$	RSS
0	10^{-5}	0.976	0.193	$2.43 \cdot 10^{-7}$	$2.33 \cdot 10^{-6}$	$3.07 \cdot 10^{-10}$
0	10^{-4}	0.976	0.192	$2.43 \cdot 10^{-7}$	$2.43 \cdot 10^{-6}$	$3.09 \cdot 10^{-10}$
0	10^{-3}	0.976	0.192	$2.43 \cdot 10^{-7}$	$2.35 \cdot 10^{-6}$	$3.09 \cdot 10^{-10}$
0	0.01	0.976	0.194	$2.44 \cdot 10^{-7}$	$2.20 \cdot 10^{-6}$	$3.16 \cdot 10^{-10}$
0	0.1	0.664	0.917	$2.17 \cdot 10^{-8}$	$1.57 \cdot 10^{-4}$	$1.08 \cdot 10^{-7}$
0	0.2	0.969	0.625	$2.14 \cdot 10^{-8}$	$3.95 \cdot 10^{-5}$	$6.88 \cdot 10^{-8}$
0.01	10^{-4}	0.976	0.193	$2.49 \cdot 10^{-7}$	$2.22 \cdot 10^{-6}$	$3.09 \cdot 10^{-10}$
0.05	10^{-4}	0.975	0.212	$2.60 \cdot 10^{-7}$	$8.20 \cdot 10^{-7}$	$3.94 \cdot 10^{-10}$
0.1	10^{-4}	0.976	0.192	$2.30 \cdot 10^{-7}$	$1.55 \cdot 10^{-6}$	$3.09 \cdot 10^{-10}$

Table 5: *Parameter estimates $\hat{\psi}_{LS}$ (transformed back into the natural parametrization) for several choices of γ_l and δ. For all cases we fixed $n = 2$ and $N_0 = 10^{10}$.*

data sets, as shown in Figure 17 and Table 6. The clear evolution of the hazard curves between the 1880s cohorts and the 1920s cohorts must be due to an increased exposure to risk factors, presumably smoking. These changes are reflected in the growing estimated proportions of high risk individuals, $1-\hat{\pi}_l$. A similar evolution of the other parameters is less clear when we consider both sexes. But the growth rates $\hat{\gamma}_u$ seem to counteract the mutation rates $\hat{\nu}$, $\hat{\mu}$. The main feature that determines the estimates is the peak in the observed hazard. Therefore, the 1920s cohorts do not contain enough information at high ages to obtain reliable $\hat{\psi}_{\text{LS}}$ values. At best, we can get an indication for the possible continuation of a trend observed in the 1880s to 1900s cohorts.

In order to assess the accuracy of the 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 two-dimensional parameter planes spanned by the four parameter axis. This reveals the strong dependencies between the different parameters. The asymptotic confidence region we get for the males 1880s cohort is shown in Figure 18. Again

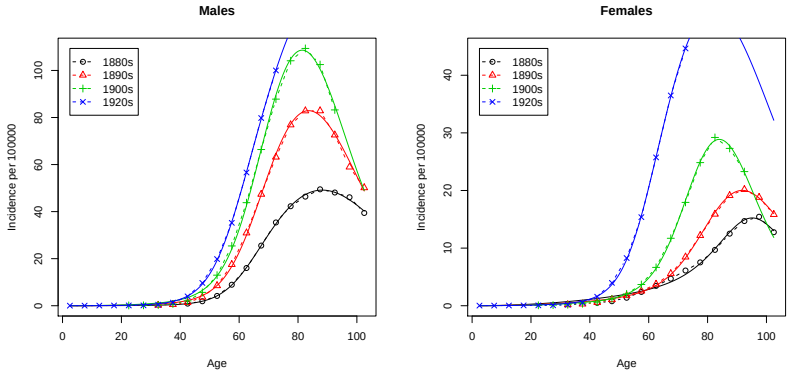

Figure 17: *Two component mixture model with frailty γ fitted to the lung cancer incidence data.*

we see the negative relationship between γ_u and the mutation rates ν, μ. The set C can be used to construct a confidence band for the hazard curve itself. Figure 19 shows all the incidence functions corresponding to parameter values contained in C. The confidence band we get reflects basically the variability present in the raw hazard estimates $\hat{\lambda}_i$.

Cohort	Males				Females			
	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{[10^{-6}]}$	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{[10^{-6}]}$
1880s	0.976	0.192	2.4	2.4	0.998	0.128	8.5	0.9
1890s	0.968	0.172	3.2	4.5	0.995	0.144	4.3	2.4
1900s	0.962	0.166	3.6	5.5	0.992	0.166	4.8	1.5
1920s	0.920	0.188	1.9	7.6	0.977	0.199	2.7	3.3

Table 6: *Parameter estimates corresponding to the curves in Figure 17. The fixed parameters are $N_0 = 10^{10}$, $\delta = 0$, $\gamma_l = 10^{-4}$ and $n = 2$.*

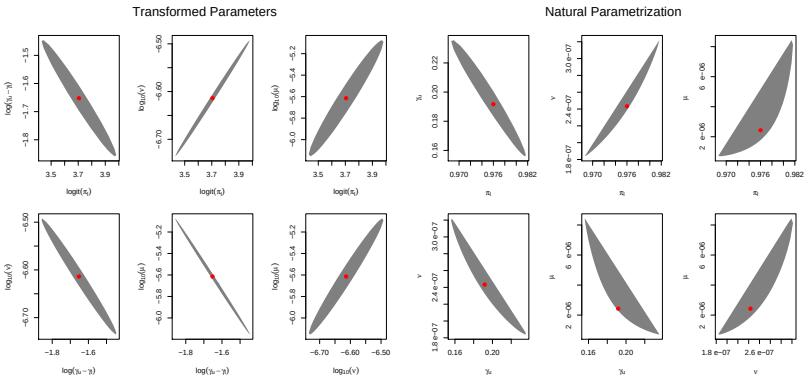

Figure 18: *Projections of an asymptotic* 95% *confidence region for* ψ .

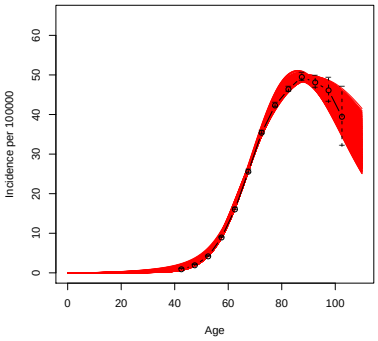

Figure 19: *Asymptotic* 95% *confidence band for the hazard function for lung cancer incidence for the males 1880s cohort. The black line gives the raw hazard estimates* $\hat{\lambda}_i$ *along with the corresponding asymptotic* 95% *confidence intervals.*

Cohort	Males				Females			
	$\hat{\pi}_l$	$\hat{\gamma}$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{u\,[10^{-5}]}$	$\hat{\pi}_l$	$\hat{\gamma}$	$\hat{\nu}_{[10^{-6}]}$	$\hat{\mu}_{u\,[10^{-6}]}$
1880s	0.977	0.150	2.7	1.4	0.995	0.074	2.4	3.0
1890s	0.967	0.152	3.4	1.1	0.993	0.081	2.5	2.2
1900s	0.960	0.146	3.8	1.3	0.991	0.138	0.5	5.5
1920s	0.932	0.174	2.2	1.2	0.981	0.172	0.3	8.3

Table 7: *Parameter estimates corresponding to the curves in Figure 20. The fixed parameters are* $N_0 = 10^{10}$, $\delta = 0$, $\mu_l = 10^{-10}$ *and* $n = 2$.

Random Rate of Malignant Transformation

Up to now, we have modelled heterogeneity in promotion only via different growth rates of initiated cells. But we can also imagine promoters that increase mutation rates. In this case, we consider μ be random with states μ_l and μ_u . This leads to a model with survivor function

$$S(t) = \pi_l S(t|\mu_l) + (1 - \pi_l) S(t|\mu_u)\,, \tag{17}$$

where $0 < \mu_l < \mu_u$. We assume n, N_0, ν, δ and γ to be equal for both components, $S(t|\mu_l)$ and $S(t|\mu_u)$, and for numerical optimization we use the parametrization

$$\tilde{\pi}_l = \text{logit}(\pi_l)\,,\ \tilde{\gamma} = \log(\gamma)\,,\ \tilde{\nu} = \log_{10}(\nu)\,,\ \tilde{\mu} = \text{logit}\left(\frac{\log_{10}(\mu_u)}{\log_{10}(\mu_l)}\right)\,.$$

Table 7 and Figure 20 summarize the estimates we get in this case. Model (17) can reproduce the lung cancer incidence data very accurately. This is a typical feature of multistage modeling: many different models fit the data equally well. But the biological interpretation is not the same for the two models (15) and (17). Therefore, model selection must be based on biological considerations and cannot be achieved by statistical means alone. Note that we get good fits only as long as μ_l is small enough. So we again have a large group that is virtually immune, and a small minority at high risk.

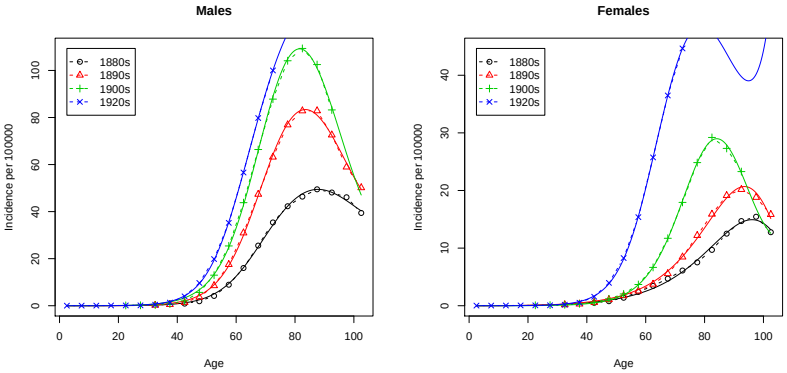

Figure 20: *Two component mixture model with frailty μ fitted to the lung cancer incidence data.*

Random Number of Mutations for Initiation

Finally, it is natural to consider a population in which a part has already inherited some mutations needed for initiation of normal stem cells. The resulting population survivor function is

$$S(t) = \pi_1 S(t|n_1) + (1 - \pi_1) S(t|n_2)\,, \tag{18}$$

with $1 \leq n_1 < n_2$. In this case, we will estimate $\tilde{\pi}_1$, $\tilde{\gamma}$, $\tilde{\nu}$ and $\tilde{\mu} = \log_{10}(\mu)$ for several fixed combinations (n_1, n_2). The results are summarized in Table 8. As expected, the number of mutations needed for initiation strongly influences the estimates $\hat{\nu}$. But also $\hat{\gamma}$ and $\hat{\pi}_1$ are affected, which is more surprising. This means that for a given cohort, the estimates $(\hat{\pi}_1, \hat{\gamma}, \hat{\nu}, \hat{\mu})|(n_1, n_2)$ differ markedly. Only some of these combinations might biologically be reasonable.

Competing Related Risks

Recall the approach proposed by Morgenthaler et al. (2004) to introduce population heterogeneity. These authors introduce two new pop-

Cohort	n_1	n_2	$\hat{\pi}_1$	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\mu}$	RSS
males							
1880s	1	2	0.038	0.133	$2.0 \cdot 10^{-11}$	$9.1 \cdot 10^{-6}$	$1.2 \cdot 10^{-9}$
1890s	1	2	0.039	0.127	$4.7 \cdot 10^{-11}$	$9.5 \cdot 10^{-6}$	$2.6 \cdot 10^{-9}$
1900s	1	2	0.046	0.129	$5.6 \cdot 10^{-11}$	$8.7 \cdot 10^{-6}$	$3.0 \cdot 10^{-9}$
1920s	1	2	0.208	0.149	$6.6 \cdot 10^{-12}$	$9.6 \cdot 10^{-6}$	$3.6 \cdot 10^{-10}$
females							
1880s	1	2	0.006	0.075	$1.1 \cdot 10^{-9}$	$3.5 \cdot 10^{-6}$	$1.7 \cdot 10^{-10}$
1890s	1	2	0.008	0.098	$1.1 \cdot 10^{-10}$	$1.0 \cdot 10^{-5}$	$2.3 \cdot 10^{-11}$
1900s	1	2	0.010	0.133	$8.8 \cdot 10^{-11}$	$2.9 \cdot 10^{-6}$	$5.9 \cdot 10^{-11}$
1920s	1	2	0.189	0.163	$3.0 \cdot 10^{-12}$	$5.2 \cdot 10^{-6}$	$2.3 \cdot 10^{-11}$
males							
1880s	2	3	0.025	0.184	$7.6 \cdot 10^{-7}$	$3.7 \cdot 10^{-6}$	$3.0 \cdot 10^{-10}$
1890s	2	3	0.034	0.161	$9.9 \cdot 10^{-7}$	$8.2 \cdot 10^{-6}$	$1.6 \cdot 10^{-9}$
1900s	2	3	0.040	0.147	$1.2 \cdot 10^{-6}$	$1.3 \cdot 10^{-5}$	$1.0 \cdot 10^{-9}$
1920s	2	3	0.069	0.174	$7.0 \cdot 10^{-7}$	$1.2 \cdot 10^{-5}$	$9.2 \cdot 10^{-11}$
females							
1880s	2	3	0.005	0.076	$4.0 \cdot 10^{-6}$	$1.0 \cdot 10^{-5}$	$2.2 \cdot 10^{-10}$
1890s	2	3	0.007	0.098	$1.6 \cdot 10^{-6}$	$2.2 \cdot 10^{-5}$	$3.1 \cdot 10^{-11}$
1900s	2	3	0.009	0.138	$1.6 \cdot 10^{-6}$	$5.8 \cdot 10^{-6}$	$9.3 \cdot 10^{-11}$
1920s	2	3	0.019	0.172	$1.1 \cdot 10^{-6}$	$8.3 \cdot 10^{-6}$	$1.6 \cdot 10^{-11}$
males							
1880s	3	4	0.017	0.333	$2.2 \cdot 10^{-5}$	$8.9 \cdot 10^{-8}$	$7.5 \cdot 10^{-10}$
1890s	3	4	0.026	0.224	$2.6 \cdot 10^{-5}$	$3.7 \cdot 10^{-6}$	$2.1 \cdot 10^{-9}$
1900s	3	4	0.032	0.173	$2.9 \cdot 10^{-5}$	$2.0 \cdot 10^{-5}$	$8.0 \cdot 10^{-10}$
1920s	3	4	0.043	0.225	$2.3 \cdot 10^{-5}$	$1.1 \cdot 10^{-5}$	$9.6 \cdot 10^{-11}$
females							
1880s	3	5	0.005	0.070	$9.7 \cdot 10^{-5}$	$5.7 \cdot 10^{-6}$	$2.0 \cdot 10^{-10}$
1890s	3	5	0.008	0.375	$2.1 \cdot 10^{-5}$	$1.4 \cdot 10^{-9}$	$8.2 \cdot 10^{-10}$
1900s	3	5	0.009	0.147	$3.4 \cdot 10^{-5}$	$1.4 \cdot 10^{-5}$	$1.5 \cdot 10^{-10}$
1920s	3	5	0.014	0.180	$3.2 \cdot 10^{-5}$	$1.9 \cdot 10^{-5}$	$1.3 \cdot 10^{-11}$

Table 8: *Model (18) applied to the lung cancer data for several fixed* (n_1, n_2) *. Note the surprisingly strong influence of the choice of* (n_1, n_2) *on* $\hat{\pi}_1$ *for the males cohorts. No good fits were obtained for the females cohorts for* $(n_1, n_2) = (3, 4)$*, but only for* $(n_1, n_2) = (3, 5)$*, which seems biologically unreasonable. In all shown cases we set* $N_0 = 10^9$ *and* $\delta = 0$ *.*

ulation parameters into the multistage model. The fraction at risk, F, quantifies the proportion of susceptibles. And the fraction of deaths due to the cancer among all deaths due to either cancer or related causes, f, models the behavior of competing but related risks. The observable hazard rate then becomes

$$h(t) \times \frac{F}{F + (1-F)\exp\left((1/f)\int_0^t h(u)\,\mathrm{d}u\right)}, \tag{19}$$

where $h(t)$ is the hazard function of the multistage carcinogenesis model.

Similarly to the previous sections, we fit model (19) to the lung cancer data from the males 1880s cohort using analytic graduation. Table 9 gives the numerical results for a list of several fixed f. We can see that completely different combinations (f, F) can lead to equivalent models in terms of goodness of fit. Nevertheless, the product $f \cdot \hat{F}$ is quite stable. Note that $f \cdot F$ corresponds to the proportion of the population that will actually die from the cancer. This corresponds to the individuals at high risk in our finite mixture models, measured for example by $\pi_u = 1 - \pi_l$ in the γ-frailty model (15).

f	$\hat{F}$	$\hat{\gamma}$	$\hat{\nu}$	$\hat{\mu}$	RSS
0.1	0.220	0.191	$7.9 \cdot 10^{-8}$	$2.9 \cdot 10^{-6}$	$2.8 \cdot 10^{-10}$
0.2	0.118	0.187	$1.1 \cdot 10^{-7}$	$3.3 \cdot 10^{-6}$	$2.9 \cdot 10^{-10}$
0.5	0.049	0.185	$1.7 \cdot 10^{-7}$	$3.5 \cdot 10^{-6}$	$3.0 \cdot 10^{-10}$
1.0	0.025	0.185	$2.4 \cdot 10^{-7}$	$3.6 \cdot 10^{-6}$	$3.0 \cdot 10^{-10}$

Table 9: *Parameter estimates for model (19) for several fixed f. The remaining parameters were set to $N_0 = 10^{10}$, $\delta = 0$, $n = 2$.*

The mixture models we presented in the previous sections did not consider competing related risks. The corresponding fraction at risk model is the one with $f = 1$. Note that in such a model there still are competing risk factors. But these factors are supposed to act in a homogeneous manner on the whole population and therefore do not alter the observable hazard. The deaths caused by such independent factors correspond in fact to the censored individuals.

Similarly to Morgenthaler et al. (2004), we can introduce related competing risks into the γ-frailty model. We make the same simplifying assumption of constant fraction of deaths due to cancer among all deaths due to either cancer or competing related causes,

$$\frac{h_{\text{pop}}(t)}{h_{\text{pop}}(t) + h_{\text{rel}}(t)} = f\,,$$

where $h_{\text{pop}}(t)$ is the population hazard function of model (15), and $h_{\text{rel}}(t)$ the hazard function due to competing related risks. Then, the observable hazard rate we fit to our data is

$$h_{\text{obs}}(t) = h_{\text{pop}}(t) \cdot f\,. \tag{20}$$

Table 10 gives the numerical results we get for several fixed f. We find again the almost constant fraction that will actually die from the cancer, which is in this model

$$f \cdot (1 - \hat{\pi}_l) \approx 0.024\,.$$

Note that only the product $f \cdot \pi_l$ is identifiable, but not the pair $(f,\ \pi_l)$.

f	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}$	$\hat{\mu}$	RSS
0.1	0.781	0.192	$2.5 \cdot 10^{-7}$	$2.8 \cdot 10^{-6}$	$2.8 \cdot 10^{-10}$
0.2	0.883	0.189	$2.5 \cdot 10^{-7}$	$3.0 \cdot 10^{-6}$	$2.9 \cdot 10^{-10}$
0.5	0.952	0.190	$2.4 \cdot 10^{-7}$	$2.8 \cdot 10^{-6}$	$3.0 \cdot 10^{-10}$
1.0	0.976	0.192	$2.4 \cdot 10^{-7}$	$2.4 \cdot 10^{-6}$	$3.1 \cdot 10^{-10}$

Table 10: *Parameter estimates for model (20) for several fixed f. The other parameters were set to $N_0 = 10^{10}$, $\delta = 0$, $\gamma_l = 10^{-4}$, $n = 2$.*

Discussion

The applications above show that many different models can fit cancer incidence data equally well. We note, however, some common features of the different models fitted to our data. For every two component

mixture model we used, we got satisfactory results only when we allowed for two clearly separated population subgroups with a low risk group that runs a risk close to zero. This immune sub-group is estimated to be the large majority of the population. For example in the males 1880s cohort we estimate that less than 2.5% are at high cancer risk. Also, the models incorporating the concept of a fraction at risk lead to the same conclusion.

The γ-frailty and μ-frailty models are such that only the parameters involved in promotion vary. This would suggest a process where initiated cells are created in all individuals (and following the same dynamics), but only in a small subgroup of the population promotion happens. This is consistent with the fact that epigenetic stimuli are necessary for the last step in malignant transformation, and such factors might be present just in a fraction of the population.

Another observation concerns the proportion of the high risk group among all individuals. The sharp increase of maximal incidence from the 1880s cohorts to the 1920s cohorts is reflected in the marked growth of $\hat{\pi}_u$.

The different models are in close agreement for the first three cohorts and differ only in the 1920s cohort, where not enough data at high ages is available. In general, the biological parameters are in good agreement between the models.

5.2.3 Counting Process Framework

The estimates we used in the previous section were all based on standard least squares. Their derivation and their asymptotic properties being a direct application of analytic graduation introduced in Section 1.3.4. A more recent and very elegant way to derive such asymptotic properties uses the counting process framework described in Section 1.3.2. The data we have at hand can be seen as arising from a counting process in discrete time, and we will use this theory to show how to derive asymptotic results in our case. We will use the same notations as before. In particular, (R_i, O_i) denote the population at risk and the observed failures at time point t_i , and $C_i = R_i - R_{i+1} - O_i$

is the number of censorings at t_i. Let $S(t)$ be the survivor function of a multistage carcinogenesis mixture model we consider, where the parameters are suppressed in the notation. Then,

$$\lambda_l = \frac{S(t_l) - S(t_{l+1})}{S(t_l)}$$

is the intensity at t_l of the corresponding failure model in discrete time.

The counting process notation can be introduced in the following way. Let T_i and X_i be failure time (i.e. time of first malignant tumor cell) and censoring time for subject i, where (T_i) and (X_i) are supposed to be independent sequences of i.i.d. random variables. We denote $\{Y_i(t), t \geq 0\}$ the individual's at risk process,

$$Y_i(t) = \mathcal{I}(T_i \geq t, X_i \geq t)\,,$$

and the process $\{N_i(t), t \geq 0\}$ counts the observed number of events up to time t for individual i,

$$N_i(t) = \mathcal{I}(T_i \leq t, T_i \leq X_i)\,.$$

Note that if $T_i = X_i$, then we count the event as an observed failure. Let n be the total number of subjects in the study, that is $n = r_0$. We observe the aggregated processes

$$N_{\cdot}(t) = \sum_{i=1}^{n} N_i(t) = \sum_{t_l \leq t} O_l\,,$$

$$Y_{\cdot}(t) = \sum_{i=1}^{n} Y_i(t) = \sum_{t_l \geq t} (O_l + C_l)\,.$$

We write the history of the process

$$\mathcal{F}_t = \sigma\{N_i(u),\ Y_i(u^+),\ i = 1, \ldots n,\ 0 \leq u \leq t\}\,,\quad t > 0\,.$$

Then, we have

$$\mathrm{P}(\,\mathrm{d}N_i(t) = 1 \,|\, \mathcal{F}_{t^-}) = Y_i(t)\,\mathrm{d}\Lambda(t)\,,$$

where the intensity process is $\Lambda(t)$ given by the step function

$$\Lambda(t) = \sum_{t_l \le t} \lambda_l \,.$$

Note that this deterministic process does not depend on individual i because the model does not include covariates.

Least Squares

In order to estimate the biological parameters, we minimize with respect to ψ the sum of squared differences between observed and expected hazard. Using the above discrete time model, we must minimize

$$\mathrm{RSS}(\psi) = \sum_{l=0}^{N} \left(\frac{O_l}{R_l \Delta t_l} - \frac{\lambda_l}{\Delta t_l} \right)^2 . \tag{21}$$

Recall that $\lambda_l = \lambda_l(\psi)$, but we will not write this dependency explicitly in order to get shorter expressions. The time points t_l are equidistant in our data set, $\Delta t_l = t_{l+1} - t_l = 5$ (years) for all l. Therefore, we can write

$$\mathrm{RSS}(\psi) \propto \sum_{l=0}^{N} \frac{1}{R_l^2} \big(O_l - R_l \lambda_l\big)^2 .$$

This leads to the estimating equations

$$\sum_{l=0}^{N} \frac{1}{R_l} \big(O_l - R_l \lambda_l\big) \frac{\partial}{\partial \psi} \lambda_l = 0 \,. \tag{22}$$

Using the counting process notation and $J(t) = \mathcal{I}(Y_{\bullet}(t) > 0)$, equation (22) is equivalent to

$$\sqrt{n} \int_0^{\tau} \frac{J(t)}{Y_{\bullet}(t)} \left(\frac{\partial}{\partial \psi} \Delta\Lambda(t) \right) \big(\mathrm{d}N_{\bullet}(t) - Y_{\bullet}(t)\, \mathrm{d}\Lambda(t) \big) = 0 \,,$$

where $\tau > a_N$ is an artificial finite end-point. In other words, we search for a solution $\hat{\psi}$ of the system

$$\int_0^{\tau} K_j(t, \psi)\ \mathrm{d}M(t, \psi) = 0 \,, \quad j = 1, \ldots, g \,,$$

where

$$K_j(t,\psi) = \sqrt{n}\frac{J(t)}{Y_{\cdot}(t)}\frac{\partial}{\partial\psi_j}\Delta\Lambda(t)\,,$$

and $\psi = (\psi_1,\ldots,\psi_g)$. Note that, with respect to $\mathcal{F}_t$, the process $K_j(t,\psi)$ is predictable and

$$M(t,\psi) = N_{\cdot}(t) - \int_0^t Y_{\cdot}(u)\,\mathrm{d}\Lambda(u)$$

is a square integrable zero mean martingale. Since $\frac{\partial}{\partial\psi_j}\Delta\Lambda(t)$ is continuous, $K_j(t,\psi)$ is bounded on the compact set $[0,\tau]$ $(t_N < \tau < \infty)$. This implies that

$$U(t,\psi) = \left(\int_0^t K_j(u,\psi)\,\mathrm{d}M(u,\psi)\right)_{j=1,\ldots,g}\,, \quad 0 \le t \le \tau\,,$$

is a square integrable multivariate martingale.

We aim at applying Rebolledo's theorem to the process $U(\tau,\psi)$. Therefore, we consider a sequence of experiments indexed by n, the total number of subjects, and introduce corresponding processes $N_{\cdot}^{(n)}(t)$, $Y_{\cdot}^{(n)}(t)$, $\Lambda^{(n)}(t)$, $U^{(n)}(t,\psi)$, etc. Recall that we assume the processes for different individuals to be independent. Let $\epsilon > 0$ be given. We define the process

$$U_{\epsilon j}^{(n)}(\tau,\psi) = \int_0^\tau K_j^{(n)}(t,\psi)\,\mathcal{I}(|K_j^{(n)}(t,\psi)| > \epsilon)\;\mathrm{d}M^{(n)}(t,\psi)\,.$$

According to Rebolledo's theorem,

$$\begin{aligned}
&\text{(i)} \qquad \langle U^{(n)}\rangle(\tau,\psi) \xrightarrow{\mathrm{P}} V \text{ as } n\to\infty\,, \text{ and}\\
&\text{(ii)} \qquad \langle U_{\epsilon j}^{(n)}\rangle(\tau,\psi) \xrightarrow{\mathrm{P}} 0 \text{ as } n\to\infty\,,\ j = 1,...,g\,,
\end{aligned}$$

together imply that

$$U^{(n)}(\tau,\psi) \xrightarrow{\mathcal{D}} \mathcal{N}(0,V)\,,$$

where V is a $g\times g$ positive semi-definite matrix. To check conditions (i) and (ii) for our situation, we first calculate[1] the predictable variation

[1] See Kalbfleisch and Prentice (2002) equation (5.34).

processes of $U_j^{(n)}(\tau,\psi)$,

$$\begin{aligned}\langle U_i^{(n)},\ U_j^{(n)}\rangle(\tau,\psi) &= \sum_{l=0}^{N} n\left(\frac{J^{(n)}(t_l)}{Y_\bullet^{(n)}(t_l)}\right)^2\left(\frac{\partial}{\partial\psi_i}\lambda_l\right)\left(\frac{\partial}{\partial\psi_j}\lambda_l\right)\\ &\qquad \times Y_\bullet^{(n)}(t_l)(1-\lambda_l)\lambda_l\\ &= \sum_{l=0}^{N} n\,\frac{J^{(n)}(t_l)}{Y_\bullet^{(n)}(t_l)}\left(\frac{\partial}{\partial\psi_i}\lambda_l\right)\left(\frac{\partial}{\partial\psi_j}\lambda_l\right)\\ &\qquad \times(1-\lambda_l)\lambda_l\,,\end{aligned}$$

and the predictable variation process of $U_{\epsilon j}^{(n)}(\tau,\psi)$,

$$\begin{aligned}\langle U_{\epsilon j}^{(n)}\rangle(\tau,\psi) &= \sum_{l=0}^{N} n\frac{J^{(n)}(t_l)}{Y_\bullet^{(n)}(t_l)}\left(\frac{\partial}{\partial\psi_j}\lambda_l\right)^2\\ &\qquad \times\mathcal{I}\left(|\sqrt{n}\frac{J^{(n)}(t_l)}{Y_\bullet^{(n)}(t_l)}\frac{\partial}{\partial\psi_j}\lambda_l|>\epsilon\right)(1-\lambda_l)\lambda_l\,.\end{aligned}$$

In the above expressions, the only random element is the at risk process $\{Y_\bullet^{(n)}(t_l)\,,\ l=0,\ldots,N\}$. From its definition it follows that

$$Y_\bullet^{(n)}(t_l)\sim\mathcal{B}(n,q_l)\,,$$

where $q_l=\mathrm{P}(Y_i(t_l)=1)=\mathrm{P}(T_i\geq t_l,X_i\geq t_l)$. According to the law of large numbers

$$\frac{Y_\bullet^{(n)}(t_l)}{n}\ \xrightarrow{\mathrm{P}}\ q_l\,,\ \text{as } n\to\infty\,.$$

We assume that the censoring mechanism is such that $q_l>0,\ \forall\, l$. This restriction ensures that function $g(x)=1/x$ is continuous at $q_l\,,\ \forall\, l$. Therefore also

$$\frac{n}{Y_\bullet^{(n)}(t_l)}\ \xrightarrow{\mathrm{P}}\ \frac{1}{q_l}\,,\ \text{as } n\to\infty\,,$$

and consequently

$$\frac{\sqrt{n}}{Y_{\bullet}^{(n)}(t_l)} \xrightarrow{\mathrm{P}} 0\,, \text{ as } n \to \infty\,.$$

This implies that

$$\langle U^{(n)} \rangle(\tau, \psi) \xrightarrow{\mathrm{P}} V\,,$$

where V is the $g \times g$ matrix with elements

$$V_{ij} = \sum_{l=0}^{N} \frac{1}{q_l} \left(\frac{\partial}{\partial \psi_i} \lambda_l \right) \left(\frac{\partial}{\partial \psi_j} \lambda_l \right) (1 - \lambda_l) \lambda_l\,.$$

It also follows that

$$\langle U_{\epsilon j}^{(n)} \rangle(\tau, \psi) \xrightarrow{\mathrm{P}} 0\,, \text{ as } n \to \infty\,.$$

Note that V is symmetric. It is also positive definite as shown by the following argument. Let us define

$$J = \begin{pmatrix} \frac{\partial}{\partial \psi_1} \lambda_0 & \cdots & \frac{\partial}{\partial \psi_g} \lambda_0 \\ \vdots & & \vdots \\ \frac{\partial}{\partial \psi_1} \lambda_N & \cdots & \frac{\partial}{\partial \psi_g} \lambda_N \end{pmatrix} \quad \text{and} \quad A = \operatorname{diag}\Big\{ \Big(\sqrt{\lambda_l (1 - \lambda_l)/q_l} \Big)_l \Big\}\,.$$

Then we have $V = J^T A^2 J$ and therefore for an arbitrary vector $x \in \mathbb{R}^g$,

$$x^T V x = (AJx)^T (AJx) \geq 0\,.$$

This shows that all conditions of Rebolledo's theorem are satisfied, and we can deduce that $U^{(n)}(\tau, \psi)$ converges to a multivariate normal distribution with mean zero and variance-covariance matrix V.

Our estimate $\hat{\psi}$ of the unknown biological parameters is defined via the estimating equations

$$U^{(n)}(\tau, \psi) = 0\,.$$

The asymptotic variance, $V(\psi, q_0, \ldots, q_N)$, of $U^{(n)}(\tau, \psi)$ depends on the unknown probabilities

$$q_l = \mathrm{P}(\text{to be at risk at time point } t_l)\,.$$

These values can be estimated by the observed proportions at risk,

$$\hat{q}_l = \frac{R_l}{R_0}\,,\ l = 0, \ldots, N\,.$$

From this we get the estimate

$$\hat{V} = V(\hat{\psi}, \hat{q}_0, \ldots, \hat{q}_N)\,,$$

which finally leads to

$$U^{(n)}(\tau, \hat{\psi}) \stackrel{\cdot}{\sim} \mathcal{N}(0, \hat{V})\,.$$

One can use this result to obtain an asymptotic confidence region for ψ. Note also that it is straightforward to generalize this argument to the weighted least squares estimate.

Modified Minimum Chi-Square

In the analytic graduation literature the use of a modified minimum chi-square approach instead of the estimating equations (21) is sometimes advocated. The idea is to estimate the variance of the raw hazard estimates, and to weight the sum of squares by the inverse of these variances. By maximum likelihood theory one can get

$$\widehat{\mathrm{Var}}(\hat{\lambda}_l) = \widehat{\mathrm{Var}}(\frac{O_l}{R_l \Delta t_l}) = \frac{O_l}{(R_l \Delta t_l)^2}\,.$$

The term *modified* is used since one uses variance estimates; see Hoem (1976) for details.

If we apply this idea to our discretized count data, we must minimize

$$\widetilde{RSS}(\psi) = \sum_{l=0}^{N} \frac{\left(\frac{O_l}{R_l \Delta t_l} - \frac{\lambda_l}{\Delta t_l}\right)^2}{O_l/(R_l \Delta t_l)^2}\,.$$

Using the same arguments as before, we get the estimating equations

$$\sum_{l=0}^{N} \frac{R_l}{O_l} \left(\frac{\partial}{\partial \psi} \lambda_l\right) (O_l - R_l \lambda_l) = 0\,.$$

In terms of our counting processes, we get the score process

$$\tilde{U}^{(n)}(\psi,\tau) = \int_0^\tau \frac{Y_\bullet^{(n)}(t)}{\sqrt{n}\Delta N_\bullet^{(n)}(t)} \left(\frac{\partial}{\partial\psi}\Delta\Lambda_l(t)\right) \mathrm{d}M_\bullet^{(n)}(t,\psi)\,,$$

where $M_\bullet^{(n)}(t,\psi)$ is the martingale defined previously. Note that we have introduced the normalizing factor $1/\sqrt{n}$. Now we get

$$\begin{aligned}\langle\tilde{U}_i^{(n)},\tilde{U}_j^{(n)}\rangle(\psi,\tau) &= \sum_{l=0}^{N} \frac{Y_\bullet^{(n)}(t_l)}{n} \left(\frac{Y_\bullet^{(n)}(t_l)}{O_l^{(n)}}\right)^2 \left(\frac{\partial}{\partial\psi_i}\lambda_l\right)\left(\frac{\partial}{\partial\psi_j}\lambda_l\right) \\ &\quad \times(1-\lambda_l)\lambda_l\,.\end{aligned}$$

Convergence of this processes can be established noting that

$$\begin{aligned}\frac{O_l^{(n)}}{Y_\bullet^{(n)}(t_l)} &\xrightarrow{\mathrm{P}} \lambda_l\,, \\ \frac{Y_\bullet^{(n)}(t_l)}{n} &\xrightarrow{\mathrm{P}} q_l\,,\end{aligned}$$

as $n\to\infty$. Finally, we get

$$\langle\tilde{U}_i^{(n)},\tilde{U}_j^{(n)}\rangle(\tau,\psi) \xrightarrow{\mathrm{P}} \tilde{V} = J^T\tilde{A}^2J\,,$$

where J is the same matrix as in the least squares case and

$$\tilde{A} = \mathrm{diag}\Big\{\left(\sqrt{q_l(1-\lambda_l)/\lambda_l}\right)_l\Big\}\,.$$

Again, we can apply Rebolledo's theorem to show convergence of the score process $\tilde{U}^{(n)}(\tau,\psi)$ to a multivariate normal with covariance matrix $\tilde{V}$.

5.3 Finite Mixtures and Colon Cancer Data

Colon cancer is among the most frequently studied cancer types, and a number of genes involved in its pathogenesis have been identified. Several mechanisms can lead to colon cancer. Nevertheless, the same signalling pathways seem to be affected. In particular, a sequence of discrete events appears to cause colon cancer, which means that a multistage process applies. See Luebeck and Moolgavkar (2005) for more biological details and for a recent application of the multistage model to colon cancer data. According to these authors, initiation is thought to be the result of mutation in both copies of the adenomatous polyposis coli (APC) gene, and therefore we assume $n = 2$. One further rare event is needed for promotion. Finally, the number of stem cells can be estimated by the number of crypts in the colon, about 10^8, multiplied by the number of stem cells per crypt. Assuming ten stem cells per crypt we therefore get $N_0 \approx 10^9$.

In the previous section, we applied several finite mixture models to lung cancer data. We can perform a similar analysis with the colon cancer data given in Table 4. Recall the corresponding raw hazard estimates shown in Figure 11. Note that for the 1920s cohort incidence data is available only up to the age interval 70-74 years. This is not enough data to observe the levelling-off of the hazard curve and its decrease at high ages. It thus makes no sense to fit our mixture models to this data set. The 1900s group also does not contain sufficient data to give reliable estimates, but in some cases the group is useful to indicate and confirm trends. We will, therefore, focus on the 1880s and the 1890s cohorts in the sequel.

Random Number of Mutations for Initiation

Let us start with model (18), which splits the population into a first group needing n_1 mutations for initiation and a second group with n_2 such mutations. In the case of colon cancer it is natural to set $n_1 = 1$ and $n_2 = 2$. The group having already inherited one mutation in the APC gene suffers from a syndrome called familial adenomatous

Cohort	Males				Females			
	$\hat{\pi}_1$	$\hat{\gamma}$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{[10^{-8}]}$	$\hat{\pi}_1$	$\hat{\gamma}$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{[10^{-8}]}$
1880s	0.025	0.095	1.5	0.9	0.022	0.093	1.2	1.4
1890s	0.030	0.117	1.8	0.1	0.024	0.105	1.4	0.4
1900s	0.038	0.099	0.4	2.5	0.043	0.077	0.1	15.8

Table 11: *Parameter estimates by least squares (via analytic graduation) corresponding to the curves in Figure 21. The fixed parameters are* $N_0 = 10^9$, $\delta = 0$, $n_1 = 1$, *and* $n_2 = 2$.

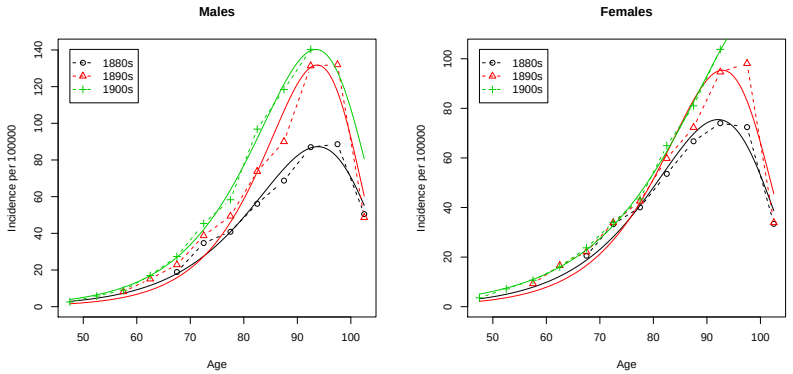

Figure 21: *Two component mixture model (18) fitted to the colon cancer incidence data by least squares (via analytic graduation).*

polyposis (FAP). These individuals usually develop multiple benign polyps in the colon, which then give rise to cancers later on in life.

We first determine the least squares estimates using analytic graduation. The numeric results are given in Table 11, and the corresponding hazard curves are illustrated in Figure 21. The model fits the data from the 1880s cohorts reasonably well. The estimated models for the 1890s cohorts, however, are less satisfactory. The very narrow peak at old ages of the observed hazard dominates the estimates. The model is not able to explain both the slight onset of cases between the ages fifty to

Cohort	Males				Females			
	$\hat{\pi}_1$	$\hat{\gamma}$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{[10^{-8}]}$	$\hat{\pi}_1$	$\hat{\gamma}$	$\hat{\nu}_{[10^{-7}]}$	$\hat{\mu}_{[10^{-8}]}$
1880s	0.026	0.085	1.3	1.9	0.023	0.082	2.5	1.3
1890s	0.033	0.095	0.6	2.0	0.028	0.084	0.9	2.9

Table 12: *Parameter estimates obtained via modified minimum chi-square graduation. The fixed parameters are* $N_0 = 10^9$, $\delta = 0$, $n_1 = 1$, *and* $n_2 = 2$.

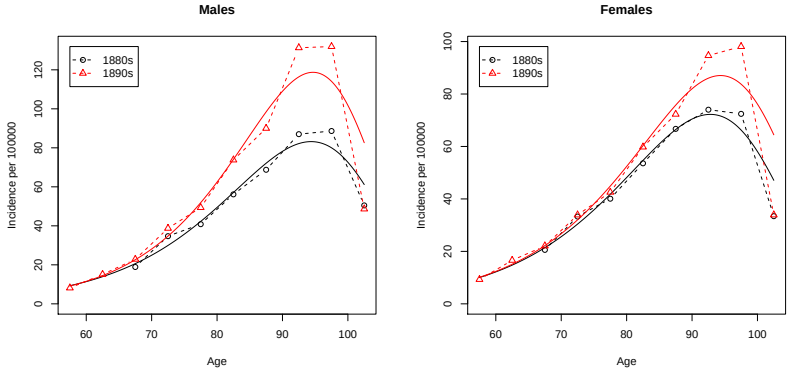

Figure 22: *Two component mixture model (18) fitted to the colon cancer incidence data by modified minimum chi-square graduation.*

eighty and the sharp peak between ages eighty and a hundred years. Since we fit by least squares, the estimate tries to avoid a large deviation at the last data points and accepts in turn many small errors at the beginning where the observed incidence is low. But the accuracy of the raw hazard estimate at 100-105 years is questionable, while we have more confidence in the first few data points. Note that the estimates are quite sensitive to the starting values chosen. In many cases the optimization scheme converges to local maxima with unacceptably poor fit. Finally, as stated above, the 1900s cohort does not contain enough information to yield trustworthy estimates. We will, therefore, carry on with only the 1880s and the 1890s cohort.

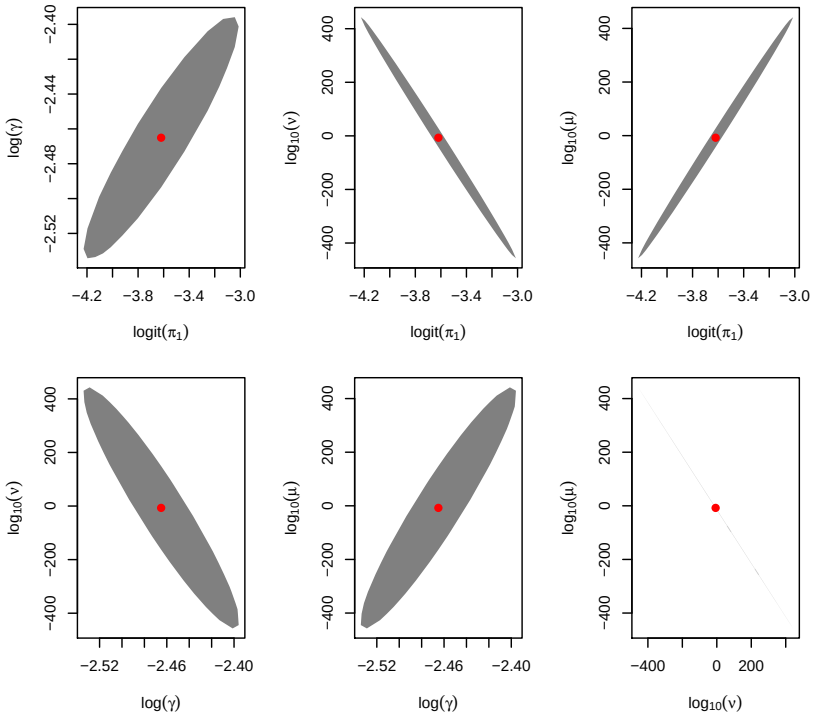

Figure 23: *Projections of an asymptotic* 95% *confidence region for parameter* ψ *.*

The above considerations suggest to downweight the influence of the raw hazard estimates at high ages, for example by estimating ψ by the modified minimum chi-square approach described previously. This means $\hat{\psi}$ is the argument that minimizes

$$Q(\psi) = (\hat{\lambda} - h(t^*; \psi))^T M (\hat{\lambda} - h(t^*; \psi)) \,,$$

where $M = \text{diag}\{\left(O_l/(R_l\Delta t_l)^2\right)^{-1};\ l = 0,\ldots,N\}$. The estimates we obtain now are given in Table 12 and Figure 22. We see that the estimates can explain the raw hazard curves very accurately up to about age eighty-five. The observed peaks at very old ages are, however, somewhat smoothed. Figure 23 gives the confidence region for the (transformed) parameter ψ. The proportion of the high risk group, π_1, is estimated to lie within the range from 1.4% to 4.7%. The net growth rate, γ, is thought to belong to the interval $[0.079, 0.091]$, which surprisingly contains only values smaller than 0.1. But we see also that the two mutation rates, ν and μ, virtually vary from 0 to ∞. These two parameters compete, and even though theoretically they are identifiable, we can in practice estimate only one of them, conditionally on the value of the other.

Random Growth Advantage

We will also fit model (15) to the colon cancer data. This model considers γ as a random variable with two states, γ_l and γ_u. We directly apply both least squares and modified minimum chi-square graduation. The results are given in Tables 13, 14 and Figure 24. Concerning the fits, the same remarks as above apply: the least squares estimates pay too much attention to the peak of the observed hazard, while the minimum chi-square estimates capture the onset of the different incidence curves better. The estimated parameters $\hat{\nu}_{\text{LS}}$ and $\hat{\mu}_{\text{LS}}$ do not seem biologically reasonable. But the corresponding estimates obtained by minimum chi-square graduation are within sensible limits. Nevertheless, when we determine confidence regions for the estimates (based on the chi-square approach), we still get huge confidence regions for the mutation rates.

Discussion

Finally, we will briefly compare our results with the initially mentioned papers by Luebeck and Moolgavkar (2002, 2005). These authors apply the multistage model to incidence data from colorectal cancer. Their data comes from US registries and covers the years from 1973 to 1996.

Cohort	Males				Females			
	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}[10^{-4}]$	$\hat{\mu}$	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}[10^{-4}]$	$\hat{\mu}$
1880s	0.984	0.114	25	$2.9 \cdot 10^{-12}$	0.987	0.118	9.7	$1.8 \cdot 10^{-11}$
1890s	0.974	0.125	0.2	$2.5 \cdot 10^{-8}$	0.982	0.121	23	$2.3 \cdot 10^{-12}$

Table 13: *Parameter estimates for the γ-frailty model applied to the colon cancer data obtained by least squares. The fixed parameters are $N_0 = 10^9$, $\delta = 0$, $n = 2$, and $\gamma_l = 0.01$.*

Cohort	Males				Females			
	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}[10^{-5}]$	$\hat{\mu}[10^{-8}]$	$\hat{\pi}_l$	$\hat{\gamma}_u$	$\hat{\nu}[10^{-5}]$	$\hat{\mu}[10^{-8}]$
1880s	0.986	0.113	4.0	1.3	0.989	0.113	3.6	1.7
1890s	0.976	0.118	3.3	1.3	0.984	0.114	4.8	0.9

Table 14: *Parameter estimates for the γ-frailty model applied to the colon cancer data obtained by modified minimum chi-square graduation. The fixed parameters are $N_0 = 10^9$, $\delta = 0$, $n = 2$, and $\gamma_l = 0.01$.*

Their model aims at the elucidation of temporal trends. Therefore, they work with the age dependent hazard curve

$$h_{ij}(t) = b_i c_j h(t) \,,$$

where b_i accounts for birth year, c_j for calendar year, and $h(t)$ is the exact hazard curve of the multistage clonal expansion model. The indices for birth year and calender year i, j and the age t are related by $i = j - t$. The parameters are then estimated by maximum likelihood, assuming that the counts in each cell of the given contingency table follow a Poisson distribution with mean $(\text{person years})_{ij} \cdot h_{ij}(t)$. The model with largest likelihood postulates two rate limiting steps for initiation and one further rate limiting step for malignant transformation. They estimate a net proliferation rate of initiated cells of 0.13 for females and 0.15 for males. The estimated mutation rates are of the order of 10^{-6} for the initiating mutations. The uncertainty for the malignant transformation is much higher, with values in the range from 10^{-9} to 10^{-6} that might all be reasonable under certain conditions. Concerning temporal trends, calendar year was most important with a peak around

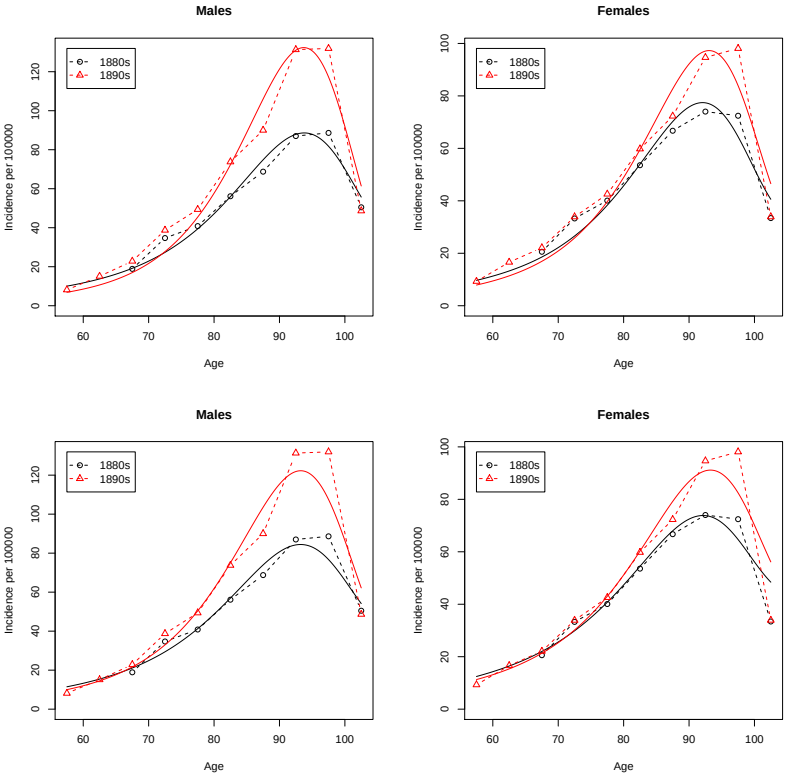

Figure 24: *Estimated hazard curves of the γ-frailty model applied to the colon cancer data. The top panels were obtained by least squares graduation, the bottom panels by minimum chi-square graduation.*

1985. Luebeck and Moolgavkar state that improved screening can be a possible explanation.

These results are consistent with our estimates from the γ-frailty model obtained by minimum chi-square graduation, reported in Table 14. Our $\hat{\gamma}_u$ is in good agreement with their values. The mutation rates

must be compared with care, since we assume $N_0 = 10^9$, while Luebeck and Moolgavkar work with $N_0 = 10^8$. But the two models seem to be in reasonable accordance. However, the scopes of the works are quite different, which explains the different modeling strategies.

5.4 Continuous Mixture Models

Up to now we have considered finite mixture models to describe the effect of heterogeneity of the population on the observed hazard curves. This implicitly assumed clearly separated population subgroups. But what would happen, if differences between individuals were less marked? We can imagine more subtle frailty effects. Some risk factors might affect a biological parameter in such a way that it varies continuously over the population. Therefore, it is natural to consider also continuous mixture models.

In Section 3.4 we introduced continuous mixtures of the multistage model. In particular, we proposed to model the net growth advantage γ by a beta distribution with support $[\gamma_l, \gamma_u]$. We will pursue this approach here. Recall the population hazard function,

$$h(t) = \int h(t|\gamma) \frac{S(t|\gamma)}{S(t)} f(\gamma)\, \mathrm{d}\gamma\,, \tag{23}$$

where the density of γ is

$$f(\gamma) = \frac{1}{B(a,b) \cdot (\gamma_u - \gamma_l)^{a+b-1}} (\gamma - \gamma_l)^{a-1} (\gamma_u - \gamma)^{b-1}\,.$$

In order to understand the effect of such a $\gamma-$frailty on $h(t)$, we calculate $h(t)$ for some choices of (a,b) for the fixed support $[\gamma_l, \gamma_u] = [0.02, 0.2]$.

Figure 25 shows the resulting population hazard curves. We have seen in Chapter 3, that monotonicity of $h(t)$ depends on the range of possible γ-values. Figure 25 illustrates another important effect: $h(t)$ has a peak if there is enough separation of the population into high risk and low risk individuals. The value of $h(t)$ at the local maximum

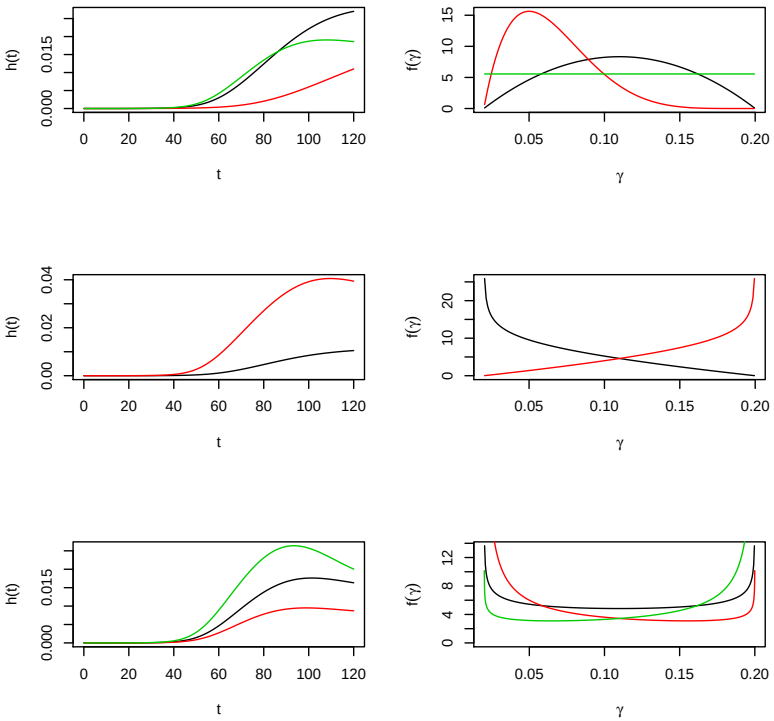

Figure 25: *Population hazard curve (23) for several shape parameters of* $f(\gamma)$. *Top row: black line,* $a = b = 2$; *red line,* $a = 2$, $b = 6$, *green line,* $a = b = 1$. *Middle row: black line,* $a = 0.8$ $b = 2$; *red line,* $a = 2$ $b = 0.8$. *Bottom row: black line,* $a = b = 0.8$; *red line,* $a = 0.4$ $b = 0.8$; *green line,* $a = 0.8$ $b = 0.4$. *The biological parameters are* $N_0 = 10^{10}$, $\delta = 0$, $\nu = 3.2 \cdot 10^{-7}$, $\mu = 10^{-5}$ *and* $n = 2$.

depends on the proportion of high risk individuals. Comparing these hazard curves to the observed hazard from lung and colon cancer, we realize that the shown models would predict too large incidences. Cancer is a rare disease, and the observed hazards count at most some hundred cases per 10^5. Figure 25 shows that realistic models could be obtained only with bathtub shaped densities $f(\gamma)$ that put most of the probability mass to the lower end of $[\gamma_l, \gamma_u]$. But such densities closely resemble two component mixtures as those studied in the previous section. This means that in our application, we do not gain further insight by working with the more complex continuous mixture models.

Discussion

The above results can be interpreted in biological terms. We found that the data suggests population heterogeneity in form of two clearly separated subgroups, rather than in form of continuous variation over some range. In order to understand this, we must keep in mind that we look at the end-point, namely cancer, of a complicated process. But biological systems are usually buffered. Small changes in the environment have no significant effect, and abrupt changes in a system may occur only after passing over some threshold value. Our population incidence data gives a global picture of the disease. There is not enough information in such data to investigate more subtle differences between individuals. But the frailty effect observed in the presented data sets can be explained by the combination of a large immune group with a small high-risk group.

Note that our conclusion is consistent with the results reported by the various approaches summarized in Chapter 3. In Morgenthaler et al. (2004) heterogeneity is modelled via the fraction at risk, F, and the fraction of deaths due to cancer, f. We have seen in Section 5.2.2 that $f \cdot F$ roughly corresponds to our proportion of high risk individuals (for example π_u in the case of the γ-frailty model). Using lung cancer incidence data one finds $f \cdot F \approx 2\%$. In Aalen and Tretli (1999) and in Moger et al. (2004) the frailty model $h_{\text{ind}}(t|Z) = Z \cdot h_0(t)$ is used, where Z is modelled with the compound Poisson distribution. This explicitly

allows for the possibility of immune individuals. These authors apply their model to Scandinavian testicular cancer incidence data. They report a proportion of susceptibles significantly lower than 0.5% for all birth cohorts before the 1950s.

CHAPTER 6

Heterogeneity

Up to now, we have discussed various ways to introduce frailty into the multistage carcinogenesis model. Our interest focused on the mathematical and statistical problems. In this final chapter, we will consider population heterogeneity itself, that is, the question of how frailty differences are created. Some examples of questions that arise with respect to heterogeneity are:

- *Sources:* which are the main factors that cause variation in cancer susceptibility? Cancer is a genetic disease, but it is also strongly related to exposure to risk factors. Which genes are involved? Which substances are carcinogenic at which doses and exposure patterns? How important are gene-environment interactions? Although a lot is known about cancer, these questions are still driving forces of a lot of biomedical research.

- *Biological mechanisms:* for known genes, which specific mechanisms at molecular, cellular or tissue level are affected, and how do these mechanisms result in heterogeneity?

- *Types of heterogeneity:* discrete groups or continuous traits, lifelong constant predisposition or time dependent relative risk, reversible or irreversible effects - which patterns of heterogeneity are observed, which can be observed?

- *Assessment:* what type of data is needed to make inference about heterogeneity? What can we infer from currently available data and how should future experiments be designed to assess heterogeneity?

All the above points pose challenges for statistical and biological modeling. We will briefly discuss some of the issues. However, we do by no means claim completeness, since a thorough treatment would be worthy of a thesis on its own. The selection of topics is somewhat arbitrary. Our aim is to illustrate further uses of statistics in carcinogenesis research. In particular, we attempt to embed the multistage model into a wider view of cancer biology, since up to now our discussion of multistage carcinogenesis has focused on the mathematical aspects.

6.1 Early Onset Cancer

The simplest possible situation are cancers caused by a single gene. We have already mentioned the prototype of such a cancer in Chapter 3: Retinoblastoma (RB), a cancer of the eye. RB is caused by a tumor suppressor gene (recessive) called $RB1$, which was the first gene to be linked to a familial cancer. The cancer is observed in a hereditary and in a sporadic form. In the former case, individuals inherit a mutated copy of the $RB1$ gene, and a single hit is enough to cause cancer. In the latter case, two somatic mutations are necessary. Mutations in $RB1$ are highly penetrant; children who inherit a mutated copy develop RB with

about 90% probability. This means that although $RB1$ mutations are physiologically recessive alleles, the syndrome is inherited dominantly. It is enough to inherit one mutated copy to be at risk. The hereditary form therefore affects mainly small children and causes a peak in the observed hazard curve at young ages. As mentioned earlier, a natural model to describe such a cancer is a two component mixture,

$$S(t) = \pi_1 S_1(t) + \pi_2 S_2(t)\,, \tag{24}$$

where $S_1(t)\,,\ S_2(t)$ are the survivor functions of a carcinogenesis model with one, and with two genetic hits respectively. The estimate $\hat{\pi}_1$ corresponds to the proportion of susceptibles to familial RB.

Simple Mendelian Models

Severe early onset diseases caused by a single gene generally show Mendelian inheritance. One can link the estimated proportion of high-risk individuals with methods from population genetics to make for instance inference on distributions of alleles. We will consider a biallelic locus, such that w denotes the wild-type allele and v the mutant allele. The genotype frequencies $p_{ww}\,,\ p_{wv}\,,\ p_{vv}$ can be expressed in terms of the allele frequencies $p_w\,,\ p_v\,.$ Under Hardy-Weinberg conditions we have $p_{ww} = p_w^2\,,\ p_{wv} = 2p_w p_v\,,\ p_{vv} = p_v^2\,.$ Mutations can perturb this equilibrium. For example in the case of a recessive mutation deleterious for reproductive fitness, only the genotypes ww and wv contribute to the next generation. For a dominant deleterious mutation solely the genotype ww is capable to reproduce. Let μ denote the rate per generation of an irreversible germ line mutation $w \to v\,.$ Then, the genotype and allele frequencies will reach an equilibrium under certain conditions:

- *dominant deleterious mutation:* since only the individuals with genotype ww reproduce, all alleles v in the population are due to new mutations. The proportion of allele v must thus be equal to the mutation rate, $p_v = \mu\,.$ Since mutation rates are usually very small, we have $p_{wv} \approx 2\mu$ and $p_{vv} \approx 0\,.$

- *recessive deleterious mutations:* homozygous variant individuals, vv, do not reproduce. An equilibrium is reached if they are replaced by new mutations. This leads to $\mu = p_{vv} = p_v^2$.

We can make similar considerations for RB and link them with the mixture model (24). But the expressions for the deleterious cases can not be applied directly, since RB is fortunately not fatal. In order to infer some information on μ, we write the genotype and the allele frequencies as a function of time, which we will count in generations. This means, we will write for generation k

$$p_{ww}(k) = p_w(k)^2\,,\ p_{wv}(k) = 2p_w(k)p_v(k)\,,\ p_{vv}(k) = p_v(k)^2\,,$$

and so on. We must also introduce ξ, the survival probability given one develops RB. Finally, let $\rho(v)$ denote the penetrance of the allele v, that is, the probability that an individual that carries the allele v will actually develop the disease. Since π_1 measures the proportion of familial RB, we can link this parameter to the frequencies of the genotypes wv and vv,

$$\pi_1 \approx [p_{wv}(k) + p_{vv}(k)]\rho(v)\,. \tag{25}$$

We will make the simplifying assumption that individuals with genotype ww will not develop RB before transmitting their alleles to the next generation. Only alleles contributed from the genotypes wv and vv are removed with a certain probability. Then, we can derive the proportions of alleles, that reach reproduction,

$$\begin{aligned}
\tilde{p}_w(k) &= p_{ww}(k) + \frac{1}{2}\{p_{wv}(k)[1-\rho(v)] + p_{wv}(k)\rho(v)\xi\} \\
&= p_{ww}(k) + \frac{1}{2}p_{wv}(k)[1-\rho(v)(1-\xi)]\,, \\
\\
\tilde{p}_v(k) &= p_{vv}(k)[1-\rho(v)] + p_{vv}(k)\rho(v)\xi + \frac{1}{2}p_{wv}(k)[1-\rho(v)(1-\xi)] \\
&= [1-\rho(v)(1-\xi)]\cdot[p_{vv}(k) + \frac{1}{2}p_{wv}(k)]\,.
\end{aligned}$$

Note that $r(t) = \tilde{p}_w(k) + \tilde{p}_v(k) < 1$. For the passage to the next generation, we must take the mutation $\mu : w \to v$ into account. The alleles, that will actually give rise to generation $k+1$, will be present in the proportions

$$\begin{aligned} \tilde{\tilde{p}}_w(k) &= \tilde{p}_w(k)(1-\mu)\,, \\ \tilde{\tilde{p}}_v(k) &= \tilde{p}_v(k) + \tilde{p}_w(k)\mu\,. \end{aligned}$$

In an infinite-alleles model, this leads to the genotype frequencies in generation $k+1$,

$$\begin{aligned} p_{ww}(k+1) &= \left(\frac{\tilde{\tilde{p}}_w(k)}{r(k)}\right)^2\,, \\ p_{wv}(k+1) &= 2\frac{\tilde{\tilde{p}}_w(k)}{r(k)}\frac{\tilde{\tilde{p}}_v(k)}{r(k)}\,, \\ p_{vv}(k+1) &= \left(\frac{\tilde{\tilde{p}}_v(k)}{r(k)}\right)^2\,. \end{aligned}$$

The normalizing factor is still the same since $\tilde{\tilde{p}}_w(k) + \tilde{\tilde{p}}_v(k) = r(k)$. If we assume that the population has reached an equilibrium state, we can derive an expression for μ by solving the equation

$$\frac{\tilde{\tilde{p}}_w(\infty)}{r(\infty)} = p_w(\infty)\,,$$

which leads to

$$\mu = \frac{\rho(v)(1-\xi)p_w(\infty)p_v(\infty)}{1-\rho(v)(1-\xi)p_v(k)}\,. \tag{26}$$

Using the information available at the website of the National Cancer Institute[1], we can get a numerical example. Namely, we get for RB the values for

- penetrance: $\rho(RB1^v) \approx 0.9$;

[1] http://cancernet.nci.nih.gov/cancertopics/types/retinoblastoma/

- survival: $\xi \approx 0.93$;
- prevalence: the cumulative hazard among children up to about five years is in the order of a few tens of cases per million, so we get roughly the order of magnitude of the high risk group, $\hat{\pi}_1 \approx 2 \cdot 10^{-5}$.

We first plug these values into equation (25) to get an estimate of $p_v(\infty)$. This value can then be used together with equation (26), which finally leads to

$$\mu \approx 7 \cdot 10^{-7} .$$

This value, a mutation rate per generation, is rather low compared to other estimates that can be found in the literature. It is not unconsistent, though, and can serve as an illustration of the use of this type of modeling.

Evolutionary Perspective

We can say that in the RB example heterogeneity is caused by a clearly identified gene. Therefore, the population is split into two subgroups, and in a mechanistic carcinogenesis model heterogeneity manifests itself through precisely determined parameters. This is an ideal but rare situation. According to Nunney (2003), it is likely that RB is the only human cancer that is regulated by a single gene. This author gives an evolutionary interpretation of this observation. Natural selection favors the suppression of pre-reproductive cancers. But the response to such selection is presumably tissue specific. The retina is a relatively small tissue, and proliferation is basically limited to the growth period of the individual. Control and repair mechanisms are expected to be more complex for cells that divide more often, at higher frequency, or that are more exposed to risk factors. One can also assume that larger animals with late reproductive age show more elaborate tumor suppressor mechanisms than smaller ones. For example, one can speculate that the number of tumor suppressor genes found in humans is larger than the number of such genes found in mice. This leads to the important

question whether conclusions from animal models can be applied to human carcinogenesis or not.

We will illustrate the necessity for more than one tumor suppressor gene as tissue size increases using a very simple model based on branching processes. A detailed discussion of such techniques can be found in Kimmel and Axelrod (2002). We will model an expanding tissue by geometric growth. That is, we consider one normal cell at time zero, $N_0 = 1$. This cell and all its descendants are assumed to divide simultaneously at every time step. Thus there are two cells at time 1, four cells at time 2, and so on. In other words, the total number of cells at time k is $N_k = 2^k$.

We will start with the simplest case of a single dominant mutation that leads to cancer. This would correspond to a (hypothetical) syndrome caused by one oncogene. We assume that during cell division a daughter cell gets the mutation with rate α, and both daughter cells act independently. Let $M_k^{(0)}$, $M_k^{(1)}$ count the number of cells without mutation (type 0 cells), and the number of cells with a mutation (type 1 cells) at time k. We assume that $(M_0^{(0)}, M_0^{(1)}) = (1, 0)$. Then, we get the transitions given in the table below.

$(M_0^{(0)}, M_0^{(1)})$	$(M_1^{(0)}, M_1^{(1)})$	with probability
(1,0)	(2,0)	$(1-\alpha)^2$
	(1,1)	$2\alpha(1-\alpha)$
	(0,2)	α^2

This means we model the expanding tissue by a multistate Galton-Watson process. The mutation is assumed to be irreversible. Therefore, a type 1 cell gives always rise to two type 1 cells at cell division. We can calculate the probability of no mutated cells at time k,

$$\begin{aligned} \mathrm{P}\left(M_k^{(1)} = 0\right) &= \mathrm{P}\left(M_k^{(1)} = 0 | M_{k-1}^{(1)} = 0\right) \mathrm{P}\left(M_{k-1}^{(1)} = 0\right) \\ &= \dots \\ &= (1-\alpha)^{2(2^k-1)}. \end{aligned}$$

Similarly, it is easy to derive the expected number of type 1 cells at

time k,

$$\begin{aligned} \mathrm{E}\left(M_k^{(1)}\right) &= \mathrm{E}\left(2M_{k-1}^{(1)} + 2\alpha(N_{k-1} - M_{k-1}^{(1)})\right) \\ &= \dots \\ &= 2^k[1-(1-\alpha)^k]\,. \end{aligned}$$

When the mutation is recessive (as is the case for tumor suppressor genes), we must introduce the cells that carry two mutations. We will say that these cells are of type 2, and count them by the process $M_k^{(2)}$. Let us note α_1 the rate of the first mutation and α_2 the rate of the second one, then we get the transitions given below.

$(M_0^{(0)}, M_0^{(1)}, M_0^{(2)})$	$(M_1^{(0)}, M_1^{(1)}, M_1^{(2)})$	with probability
(1,0,0)	(2,0,0)	$(1-\alpha_1)^2$
	(1,1,0)	$2\alpha_1(1-\alpha_1)$
	(0,2,0)	α_1^2
(0,1,0)	(0,2,0)	$(1-\alpha_2)^2$
	(0,1,1)	$2\alpha_2(1-\alpha_2)$
	(0,0,2)	α_2^2
(0,0,1)	(0,0,2)	1

In this case, it is useful to introduce the probability generating functions $g_i(s_0, s_1, s_2; k)$, $i = 0, 1, 2$, of the vector $(M_k^{(0)}, M_k^{(1)}, M_k^{(2)})$, where $g_i(s; k)$ is the function we get when the process starts from a single type i cell at time $k = 0$. Using the above table, one can derive the recursions

$$\begin{aligned} g_0(s; k+1) &= [(1-\alpha_1)g_0(s;k) + \alpha_1 g_1(s;k)]^2\,, \\ g_1(s; k+1) &= [(1-\alpha_2)g_1(s;k) + \alpha_2 g_2(s;k)]^2\,, \\ g_2(s; k+1) &= g_2(s;k)^2\,, \end{aligned}$$

with initial condition $g_i(s; 0) = s_i$, $i = 0, 1, 2$. We will note $p_i(k) = g_i(1, 1, 0; k)$ the probability of no type 2 cell at time k. The recursive equations imply that $p_2(k) = 0$, $p_1(k) = (1-\alpha_2)^{2(2^k-1)}$, and

$$p_0(k+1) = [(1-\alpha_1)p_0(k) + \alpha_1 p_1(k)]^2\,, \quad p_0(0) = 1\,.$$

The expected number of type 2 cells can be derived using

$$\mathrm{E}\left(M_k^{(2)} \,\middle|\, (M_0^{(0)}, M_0^{(1)}, M_0^{(2)}) = (1,0,0)\right) = \frac{\partial}{\partial s_2} g_0(s;k)\bigg|_{s_0=s_1=s_2=1}.$$

In order to estimate the numbers of cell divisions necessary to build a tissue under such a model, Table 15 summarizes rough estimates of the size of some organisms or tissues that can be found in the literature. The probabilities of no mutated cells as a function of tissue size for both the dominant and the recessive case are shown in Figure 26. It becomes clear that only for very small tissues, such as the retina, a single tumor suppressor gene may be enough to prevent the occurrence of cancer before reproductive age.

Species/Tissue	Estimated Size (Nb. of cells C)	Nb. of Doublings ($k = \log_2 C$)
C. elegans[2]	10^3	9.9
D. melanogaster[2]	$5{\cdot}10^6$	22.3
Retina[3]	$4{\cdot}10^6$	21.9
Colon (dividing cells)[4]	10^{10}	33.2
Human body (adult)[5]	$5{\cdot}10^{13}$	45.5

Table 15: *Estimated numbers of cells for some organisms or tissues.*

6.2 Association Studies

The study of early onset cancers shows that almost all cancers are caused by more than one gene. When we try to get an overview of the genes that are known to be related to cancer, we realize that the list of such genes is actually very long. A review of currently known

[2]See Nunney (1993).
[3]See Hethcote and Knudson (1978).
[4]See Michor et al. (2004).
[5]See Tomlinson et al. (2002).

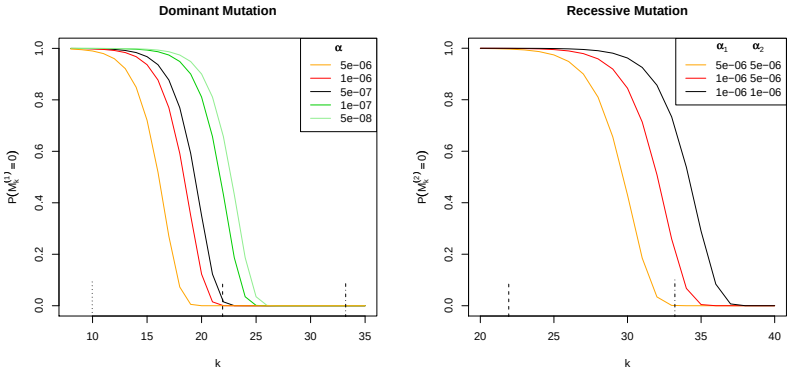

Figure 26: *Probability of no state 1 cell (dominant case), and of no state 2 cell (recessive case). The vertical lines indicate the estimated number of cells in C. elegans (dotted line), in the retina (dashed line), and in the colon.*

cancer genes can be found in Futreal et al. (2004). These authors give a list of 291 genes that have been reported to be causally implicated in oncogenesis. This means that up to now more than 1% of the total human genome has been linked to cancer. Such cancer genes are mainly related to the regulation of cell proliferation, cell differentiation, cell death, and DNA repair. But there seems also to be a strong genetic control of the amount of hormones a tissue is exposed to. So in some cases such as breast cancer susceptibility genes could also be involved in hormone pathways (see Kolonel et al. (2004)).

The identification of susceptibility genes traditionally uses family studies and population based association studies (case-control studies). Futreal et al. state that up to now most associations reported in the literature have not been confirmed in subsequent surveys. They mention false positives and publication bias as possible explanations. Population heterogeneity can be another reason for a lack of reproducibility of reported results. According to Cardon and Palmer (2003), allele frequencies vary quite markedly within and between populations - often irrespective of disease status. Furthermore, cancer susceptibility alle-

les are generally very rare. This shows that these alleles are of recent origin. This is confirmed by the observation, that there are many different susceptibility genes, and that those genes have different effects in different tissues. Moreover, there seems to be a species barrier: many cancer genes affect different organs when mutated in mice than when mutated in humans (see Vogelstein and Kinzler (2004)).

In the study of cancer causing genes not only mutations, but also regulation is of importance. Changes in phenotype - including cancer - must not necessarily be due to changes in genotype. They can also be a consequence of epigenetic regulation. For example, tumor suppressor genes might be epigenetically silenced. Epigenetic alterations presumably reflect exposure to risk factors. This shows that it is not enough to identify lists of genes that are related to carcinogenesis. The understanding of gene-gene and gene-environment interactions is the real challenge, and population heterogeneity with respect to cancer risk seems to stem from the combination of polygenic interactions with specific environmental settings.

6.3 Gene-Gene Interactions

The ability of one or more genes to alter the action of another gene is called epistasis. A simple example of epistasis is the albino gene, which can mask the effect of the gene that determines hair color. The discovery and elucidation of gene-gene interactions is a difficult task from both the biological and the statistical point of view. A review of mathematical methods used to investigate multi-locus traits can be found in Hoh and Ott (2003).

A natural idea is to try to extend single locus methods to the multi-locus case. One such approach is the use of sums of single-marker statistics. For example, one can calculate a measure of association (as the χ^2 statistics) using contingency tables for every marker. This would lead to a sequence $t_1, \ldots, t_n$. An ad hoc procedure to identify groups of interacting genes consists in combining the order statistics,

$t_{(1)}, \ldots, t_{(n)}$, into sums

$$s_k = \sum_{i=n-k+1}^{n} t_{(i)} .$$

For fixed k one can assess p-values by resampling methods. This can be repeated for several k , and one can select subsets of markers that seem to be most important. Such methods must find a trade off between selecting enough genes to hopefully capture those related to the disease and between selecting to many markers and thereby accepting a high proportion of false positives. But in any case, such procedures can only propose candidates for further investigation. Whether interaction is present and how it could work must then be studied using more sophisticated modeling. See Hoh et al. (2000) for more details.

To directly investigate interactions among genes, we should clearly prefer methods that make a joint analysis of all genes. In principle, logistic regression can be used. However, the number of possible interactions grows exponentially with the number of makers. Therefore, the number of parameters to be estimated in a full model becomes quickly too large. In practice, often only pairwise interactions can be explained.

The difficulty of parametric models to handle high-order interactions suggests the use of model free methods. Given the usually very large number of candidate genes, it seems appealing to use statistical learning theory to recognize patterns of genes related to disease outcome. We will consider one recently proposed such method in some more detail.

Polymorphism Interaction Analysis

In Goodman et al. (2006) an algorithm called Polymorphism Interaction Analysis (PIA) was proposed. It aims at the identification of interactions among Single Nucleotide Polymorphisms (SNPs) that result in an increase in colon cancer risk. PIA is a slight generalization of a method called Multifactor-Dimensionality Reduction (MDR), which was first proposed in Ritchie et al. (2001). MDR was designed to detect

high-order gene-gene (and gene-environment) interactions using genotyped case-control data. That is a set of cancer patients and a set of matched controls are chosen. Then, the genotypes at N selected SNP-sites are determined. The SNPs are defined a priori and supposed to be relevant for the disease.

The MDR algorithm is a simple classification scheme for which the mis-classification and the prediction errors are estimated via cross-validation (CV). A subset of K SNPs is selected, then the 3^K genotypes are classified as *high risk* and *low risk* according to whether

$$\frac{\text{nb. of cases } (genotype)}{\text{nb. of controls } (genotype)} \geq T$$

or not. The threshold T is fixed in advance; usually $T = 1$. For a given order of interactions, that means for a given number K, MDR chooses among the $\binom{N}{K}$ combinations the set of SNPs with the lowest CV prediction error. PIA is a generalization in the sense that it allows for two different cost functions to judge the quality of the model. The first is the Gini index (see for example Hastie et al. (2001)). The second one, called %-wrong, is simply the misclassification error of the chosen model when applied again to the whole data (as opposed to the data split into training and testing sets).

In the article MDR was proposed, the method was applied to sporadic breast cancer data. It was claimed that MDR had identified a new four locus interaction among a list of 10 candidate SNPs. The CV prediction error of the selected model was 46.7%. In Xu et al. (2005) MDR was applied to prostate cancer. Their analysis included 57 SNPs. Again a four locus model was reported to predict cancer risk best. In this case, the estimated prediction error was 37.7%.

The researchers that proposed PIA used colon cancer data. They explored all combinations up to fourth order among 94 SNPs. The reported %-wrong scores of the best first, second, third, and fourth order models were 43.7%, 40.0%, 35.3%, and 31.7%. The researchers admitted in their discussion the possibility, that a SNP combination could have been selected due to chance as a consequence of the large

number of sites considered. But they did not really account for multiplicity in their analysis. Consequently, these authors see their method as an identification procedure, which can narrow down the list of SNPs that seem to be able to explain cancer risk. They do not claim PIA to be powerful enough to reliably find gene-gene interactions, and do not advice to apply PIA as a stand alone method.

From the statistical point of view, the above described results are not satisfactory at all. The prediction errors are extremely large, and can actually be explained solely by random variation. We will illustrate this by means of a simulation. For 200 cases and 200 controls, we randomly generate genotypes. We sample for every SNP independently from the same multinomial distribution,

$$\mathcal{M}\left(1\,;\; p_w^2\,,\; 2p_w(1-p_w)\,,\; (1-p_w)^2\right)\,.$$

This means we assume a Hardy-Weinberg equilibrium for every locus with frequency of the wild-type allele p_w . We then apply the MDR/PIA classification for several numbers of candidate SNPs. Figure 27 shows the CV classification and prediction errors for second and third order interactions. We see that the prediction and classification errors drop markedly as the number of candidate SNPs or the order of interactions increase. The values obtained by our simulation are comparable to the results of the above described studies. So in our example, we find SNP combinations that discriminate cases and controls equally well as the interactions reported in the literature. Nevertheless, there is no relationship between disease status and genotypes in our simulated datasets.

Besides the problem of multiplicity, our simulation brings up another question: which part of cancer incidence can be explained by genetic predisposition? In other terms: how much is environmental, and how much is genetic? It is an appealing idea to relate cancers that cannot be explained by single genes and Mendelian patterns to more complicated gene-gene interactions. But the approach of just listing enough genes and then comparing genotype frequencies in cases and controls is too simplistic. Even when methods such as MDR/PIA

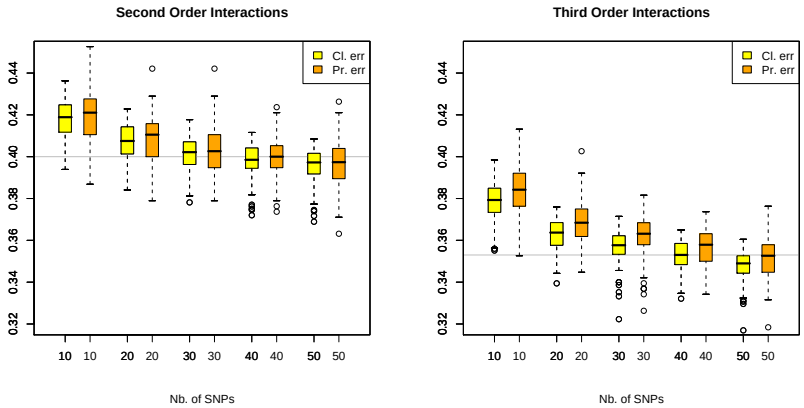

Figure 27: *Boxplots of CV classification and prediction errors from 100 simulations for several numbers of candidate SNPs. The frequency of the wild-type was* $p_w = 0.7$ *, and we split the data into 10 parts for cross-validation. These values were chosen in order to mimic the conditions of the studies, where MDR/PIA were proposed. The horizontal grey lines give the %-wrong scores of the best models found in Xu et al. (2005) using PIA for the corresponding order of interactions. We get models with comparable CV errors from about 50 SNPs on, though the above mentioned authors included 94 SNPs in their analysis.*

would account properly for the multiplicity problem, they must fail in situations, where exposure to carcinogens and lifestyle are the main risk factors. If cancer is caused predominantly by sporadic mutations, and if the impact of genetic predisposition is small, then such interaction analysis will not give meaningful results. Note that environmental agents could be considered in MDR/PIA. The inventors of MDR state that not only SNPs, but any discrete variable can be integrated in their analysis. One simply has to define exposure groups. The environmental factors are then treated just as the SNPs in the analysis.

The issue of genes and environment is crucial, and we will discuss the problem of how to disentangle genetic and environmental causes of cancer in the next section.

6.4 Familial Risk

The classical approach to estimate the heritable contribution to cancer uses family and especially twin data. For example, one can determine standardized incidence ratios (SIR) for monozygotic (MZ) and dizygotic (DZ) twins. SIR means the observed incidence rate for a given group is determined and divided by the expected incidence rate for the general population. Since MZ twins share twice as much genetic information as DZ twins, a purely additive genetic effect should lead to a ratio $\mathrm{SIR_{MZ}}/\mathrm{SIR_{DZ}}$ of about two. A ratio larger than two, however, would suggest interactions: either non-additive genetic effects or gene-environment interactions. The same considerations hold for the comparison of full siblings with half siblings. An application of this idea using Swedish twin data can be found in Ahlbom et al. (1997).

Besides the computation of relative risks, methods from quantitative genetics can be used to assess genetic and environmental contributions to cancer. Since one deals with a dichotomous outcome, generalizations of the methods for continuous variables must be used. The liability-threshold model assumes that cancer is determined by a standard normal random variable D, such that a person gets affected if D exceeds some unknown threshold. In a structural model the liability D is decomposed into several sources. For example in Czene et al. (2002) the model

$$D = G + S + F + E \tag{27}$$

is used, where G accounts for genetic effects, S for shared environmental effects, F for shared childhood environment effects, and E for unshared environmental effects. All variables are assumed to be centered and independent. Since D is standard normal, one gets the variance decomposition

$$\sigma_{\mathrm{G}}^2 + \sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{F}}^2 + \sigma_{\mathrm{E}}^2 = 1\,.$$

These variances can be estimated using the known correlation structure among family members. This means one assumes

- for two spouses: $\mathrm{Cov}(D_1, D_2) = \sigma_{\mathrm{S}}^2$;

- for full sibs: $\mathrm{Cov}(D_1, D_2) = 0.5\sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{S}} + \sigma^2_{\mathrm{F}}$;
- for half sibs: $\mathrm{Cov}(D_1, D_2) = 0.25\sigma^2_{\mathrm{G}} + \sigma^2_{\mathrm{S}} + \sigma^2_{\mathrm{F}}$.

Equating this to the corresponding estimated tetrachoric correlations, one gets a system of linear equations for the unknown variances. The tetrachoric correlation estimated from a 2×2 contingency table for a liability-threshold model is an estimate of the correlation between the underlying continuous variables D_1, D_2 . For more details see for example Sham (1998).

According to Czene et al. (2002) the factor E plays by far the most important role for most cancer sites. These authors applied model (27) to 15 common cancers. The genetic contribution to the variance exceeded 0.5 only for the thyroid. For all other sites, $\hat{\sigma}^2_{\mathrm{G}}$ was between 0.01 and 0.28 . Their estimated contributions to liability for lung and colon cancer are given in Table 16. Their results are based on the Swedish Family-Cancer Database, which contains more than 9.6 million individuals. Similar results were reported in several related articles such as Hemminki et al. (2001, 1998).

Site	$\hat{\sigma}^2_{\mathrm{G}}$	$\hat{\sigma}^2_{\mathrm{S}}$	$\hat{\sigma}^2_{\mathrm{F}}$	$\hat{\sigma}^2_{\mathrm{E}}$
Lung	0.08 (0.05-0.09)	0.09 (0.08-0.09)	0.04 (0.00-0.04)	0.79 (0.79-0.82)
Colon	0.13 (0.12-0.18)	0.12 (0.11-0.13)	0.06 (0.05-0.07)	0.69 (0.68-0.70)

Table 16: *Estimated variance contributions (and* 95% *confidence intervals) from the structural model (27) given in Czene et al. (2002).*

The structural model outlined above can of course be criticized. The assumptions of additivity, linearity and normality can be questioned. Confounding is also a major problem. It is for example not so clear what the shared childhood environment factor F really measures. The genetic contribution G only accounts for additive genetic effects. Interactions, gene-gene or gene-environment, will actually contribute to F . It does not seem feasible to make proper inference on

gene-environment interactions within this approach, since one had to observe relatives over a range of known environmental settings. Also, genetic resemblance between spouses is neglected. Therefore, Czene et al. (2002) state that the heritable contribution G should be considered as a lower bound of the importance of genetic effects.

Even if we accept the model as it is, its interpretation remains difficult. The estimates in Table 16 show for example no significant shared childhood environment effect for lung cancer. What does this mean? The role of the specific childhood environment in lung cancer is controversial. The main contribution to liability for lung cancer, the residual environmental effect E, does not exclude exposure during childhood. But it accounts solely for non-shared effects, that means individual, sporadic causes, which can happen at any point in life. Finally, the results in Table 16 do not only contain biological information, they might also be heavily influenced by the cultural and social standards of the population studied.

6.5 Conclusion

The previous section has show that unknown environmental factors, E, play a key role in cancer. Non-shared environmental effects mean sporadic, eventually unique events, and they express the inherently random part of carcinogenesis. Nevertheless, the progress in cancer therapy shows that our understanding of cancer has grown rapidly over the last decades. The short overview over some topics presented in this chapter has shown various uses of statistical methods in cancer research. The last two sections have illustrated in particular how difficult it is to extract relevant information from population based studies. It is our conviction, that a mechanistic understanding - and therefore mechanistic modeling - is crucial for a successful investigation of open questions such as the interaction between genes and between genes and environment. It is certainly interesting to work with the high throughput methods from modern molecular biology, and the needed statistical learning algorithms get more powerful. But such approaches should be

used to elucidate well defined, clear cut problems. Blind applications without detailed biological knowledge seem too optimistic. Statistics can certainly contribute its part, but it should be brought in with a modeling perspective in mind. This is supported by our impression that many of the most relevant insights into cancer originate from clinical research.

Bibliography

Odd O. Aalen. *Statistical inference for a family of counting processes.* PhD thesis, University of California, Berkley., 1975.

Odd O. Aalen. Effects of frailty in survival analysis. *Stat. Meth. Med. Res.*, 3:227–243, 1994.

Odd O. Aalen and Håkon K. Gjessing. Understanding the shape of the hazard rate: a process point of view. *Statist. Sci.*, 16(1):1–22, 2001. ISSN 0883-4237. With comments and a rejoinder by the authors.

Odd O. Aalen and Steinar Tretli. Analyzing incidence of testis cancer by means of a frailty model. *Cancer Causes and Control*, 10:285–292, 1999.

Anders Ahlbom, Paul Lichtenstein, Håkan Malmström, Maria Feychting, Kari Hemminki, and Nancy L. Pedersen. Cancer in twins: Genetic and nongenetic familial risk factors. *JNCI*, 89(4):287–293, 1997.

Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0.

P. Armitage and R. Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Brit. J. Cancer*, 8(1):1–12, 1954.

P. Armitage and R. Doll. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Br. J. Cancer*, 11: 161–169, 1957.

Richard E. Barlow and Frank Proschan. *Statistical theory of reliability and life testing.* Holt, Rinehart and Winston, Inc., New York, 1975. Probability models, International Series in Decision Processes, Series in Quantitative Methods for Decision Making.

Henry W. Block, Yulin Li, and Thomas H. Savits. Initial and final behaviour of failure rate functions for mixtures and systems. *J. Appl. Probab.*, 40(3):721–740, 2003. ISSN 0021-9002.

N. E. Breslow and N. E. Day. *Statistical Methods in Cancer Research. 1: The analysis of case-control studies.* I.A.R.C., Lyon, 1980. ISBN 92-832-0132-9.

Lon R. Cardon and Lyle J. Palmer. Population stratification and spurious allelic association. *The Lancet*, 361:598–604, February 2003.

Noel Cressie. *Statistics for Spatial Data*, chapter Modeling Growth with Random Sets, pages 776–802. Wiley, New York, 1991.

Kamila Czene, Paul Lichtenstein, and Kari Hemminki. Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *Int. J. Cancer*, 99:260–266, 2002.

R. Doll and A.B. Hill. Smoking and carcinoma of the lung. preliminary report. *British Medical Journal*, 2:739–748, 1950.

Richard Doll, Richard Peto, Jillian Boreham, and Isabelle Sutherland. Mortality in relation to smoking: 50 years' observations on male british doctors. *BMJ*, 328(1519), 2004.

Lutz Edler and Christos P. Kitsos. *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*. John Wiley & Sons, Ltd., West Sussex, England, 2005.

Lutz Edler and Annette Kopp-Schneider. Origin of the mutational origin of cancer. *International Journal of Epidemiology*, 34(5):1168–1170, 2005.

L. M. Franks and N. M. Teich, editors. *Cellular and Molecular Biology of Cancer*. Oxford University Press, third edition, 2001.

P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. A census of human cancer genes. *Nat. Rev. Cancer*, 4:177–183, March 2004.

Julie E. Goodman, Leah E. Mechanic, Brian T. Luke, Stefan Ambs, Stephen Chanock, and C. Harris Curtis. Exploring snp-snp interactions and colon cancer risk using polymorphism interaction analysis. *Int. J. Cancer*, 118:1790–1797, 2006.

John Gurland and Jayaram Sethuraman. How pooling failure data may reverse increasing failure rates. *J. Amer. Statist. Assoc.*, 90 (432):1416–1423, 1995. ISSN 0162-1459.

Leonid G. Hanin. Identification problem for stochastic models with application to carcinogenesis, cancer detection and radiation biology. *Discrete Dynamics in Nature and Society*, 7(3):177–189, 2002.

Leonid G. Hanin and Kenneth M. Boucher. Identifiability of parameters in the yakovlev-polig model of carcinogenesis. *Mathematical Biosciences*, 160:1–24, 1999.

Leonid G. Hanin and Andrej Yu Yakovlev. A nonidentifiability aspect of the two-stage model of carcinogenesis. *Risk Analysis*, 16(5):711–715, 1996.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

W. F. Heidenreich. On the parameters of the clonal expansion model. *Radiat. Environ. Biophys.*, 35:127–129, 1996.

W. F. Heidenreich, E. G. Luebeck, and S. H. Moolgavkar. Some properties of the hazard function of the two-mutation clonal expansion model. *Risk Analysis*, 17(3):391–399, 1997.

Kari Hemminki, Ingrid Lönnstedt, Pauli Vaittinen, and Paul Lichtenstein. Estimation of genetic and environmental xomponents in colorectal and lung cancer and melanoma. *Genetic Epidemiology*, 20: 107–116, 2001.

Kari Hemminki, Pauli Vaittinen, and Pentti Kyyrönen. Age-specific familial risks in common cancers of the offspring. *Int. J. Cancer*, 78: 172–175, 1998.

P. Herrero-Jimenez, A. Tomita-Mitchell, E. E. furth, S. Morgenthaler, and W. G. Thilly. 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*, 447:73–116, 2000.

Herbert W. Hethcote and Alfred G. Knudson. Model for the incidence of embryonal cancers: Appliation to retinoblastoma. *Proc. Natl. Acad. Sci. USA*, 75(5):2453–2457, 1978.

Jan M. Hoem. 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), Vol. I: Theory of statistics*, pages 569–600, Berkeley, Calif., 1972. Univ. California Press.

Jan M. Hoem. The statistical theory of demographic rates. A review of current developments. *Scand. J. Statist.*, 3(4):169–185, 1976. ISSN 0303-6898. With discussion by Niels Keiding, Hannu Kulokari, Bent Natvig, Ole Barndorff-Nielsen, Jørgen Hilden and a reply by the author.

J. Hoh, A. Wille, R. Zee, S. Cheng, R. Reynolds, K. Lindpaintner, and J. Ott. Selecting snps in two-stage analysis of disease association data: a model-free approach. *Ann. Hum. Genet.*, 64:413–417, 2000.

Josephine Hoh and Jurg Ott. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Reviews Genetics*, 4: 701–709, September 2003.

Philip Hougaard. Frailty models for survival data. *Lifetime Data Analysis*, 1:255–273, 1995.

Shizue Izumi and Megu Ohtaki. Aspects of the armitage-doll gamma frailty model for cancer incidence data. *Environmetrics*, 15:209–218, 2004.

Richard J. Jones, William H. Matusi, and B. Douglas Smith. Cancer stem cells: Are we missing the target. *J.Nat.Canc.Inst.*, 96(8):583–585, 2004.

John D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2002. ISBN 0-471-36357-X.

David G. Kendall. Birth-and-death processes, and the theory of carcinogenesis. *Biometrika*, 47:13–21, 1960.

Marek Kimmel and David E. Axelrod. *Branching Processes in Biology*. Springer, New York, 2002.

Alfred G. Knudson. Two genetic hits (more or less) to cancer. *Nature Reviews, Cancer*, 1:157–162, November 2001.

Laurence N. Kolonel, David Altshuler, and Brian E. Henderson. The multiethnic cohort study: Exploring genes, lifestyle and cancer risk. *Nature Reviews, Cancer*, 4:1–9, July 2004.

Anette Kopp-Schneider. Carcinogenesis models for risk assessment. *Statistical Methods in Medical Research*, 6:317–340, 1997.

Annette Kopp-Schneider and Christopher J. Portier. A stem cell model for carcinogenesis. *Math. Biosci.*, 120:211–232, 1994.

Annette Kopp-Schneider, Christopher J. Portier, and Claire D. Sherman. The exact formula for tumor incidence in the two-stage model. *Risk Analysis*, 14(6):1079–1080, 1994.

Jerald F. Lawless. *Statistical models and methods for lifetime data*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2003. ISBN 0-471-37215-3.

M. P. Little. 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*, 51:1278–1291, 1995.

M. P. Little and E. G. Wright. A stochastic carcinogenesis model incorporating genomic instability fitted to colon cancer data. *Mathematical Biosciences*, 183:111–134, 2003.

E. Georg Luebeck and Suresh H. Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *PNAS*, 99(23):15095–15100, 2002.

E. Georg Luebeck and Suresh H. Moolgavkar. Biological and mathematical aspects of multistage carcinogenesis. In Lutz Edler and Christos P. Kitsos, editors, *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*. Wiley, 2005.

James D. Lynch. On conditions for mixtures of increasing failure rate distributions to have an increasing failure rate. *Probab. Engrg. Inform. Sci.*, 13(1):33–36, 1999. ISSN 0269-9648.

Nicole F. Mathon and Alison C. Lloyd. Cell senescence and cancer. *Nature Reviews, Cancer*, 1:203–213, December 2001.

Franziska Michor, Yoh Iwasa, and Martin A. Nowak. Dynamics of cancer progression. *Nature Reviews, Cancer*, 4:197–205, March 2004.

Tron A. Moger, Odd O. Aalen, Tarje O. Halvorsen, Hans H. Storm, and Steinar Tretli. Frailty modelling of testicular cancer incidence using scandinavian data. *Biostatistics*, 5(1):1–14, 2004.

Suresh H Moolgavkar. Commentary: Fifty years of the multistage model: remarks in a landmark paper. *International Journal of Epidemiology*, 33(6):1182–1183, 2004.

Suresh H. Moolgavkar and A.G. Knudson. Mutation and cancer: A model for human carcinogenesis. *J. Nat. Cancer Inst.*, 66:1037–1052, 1981.

Suresh H Moolgavkar and E. Georg Luebeck. Multistage carcinogenesis and the insidence of human cancer. *Genes, Chromosomes & Cancer*, 38:302–306, 2003.

Suresh H. Moolgavkar and Georg Luebeck. Two-event model for carcinogenesis: Biological, mathematical, and statistical considerations. *Risk Analysis*, 10(2):323–341, 1990.

Suresh H. Moolgavkar and David J. Venzon. Two-event models for carcinogenesis: Incidence curves for childhood and adult tumors. *Mathematical Biosciences*, 47:55–77, 1979.

Stephan Morgenthaler, Pablo Herrero, and William G. Thilly. Multistage carcinogenesis and the fraction at risk. *Journal of Mathematical Biology*, 49(5):455–467, 2004.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7:308–313, 1965.

C. O. Nordling. A new theory on the cancer-inducing mechanism. *Brit. J. Cancer*, 7:68–72, 1953.

Leonard Nunney. Lineage selection and the evolution of multistage carcinogenesis. *Proc. R. Soc. Lond. B*, 266:493–498, 1993.

Leonard Nunney. The population genetics of multistage carcinogenesis. *Proc. R. Soc. Lond. B*, 270:1183–1191, 2003.

Ricardo Pardal, Michael F. Clarke, and Sean J. Morrison. Applying the principles of stem-cell biology to cancer. *Nature Reviews, Cancer*, 3: 895–902, December 2003.

Luigi Preziosi, editor. *Cancer Modelling and Simulation*. Chapman & Hall/CRC Mathematical Biology and Medecine Series. Chapman & Hall, London, 2003.

Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, and Jason H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, 69:138–147, 2001.

Raymond W. Ruddon. *Cancer Biology*. Oxford University Press, New York, third edition edition, 1995.

Erkki Ruoslahti. How cancer spreads. *Scientific American*, 275:42–47, September 1996.

Moshe Shaked and Fabio Spizzichino. Mixtures and monotonicity of failure rate functions. In *Advances in reliability*, volume 20 of *Handbook of Statist.*, pages 185–198. North-Holland, Amsterdam, 2001.

Pak Sham. *Statistics in Human Genetics*. Arnold, London, 1998.

Claire D. Sherman and Christopher J. Portier. The two-stage model of carcinogenesis: Overcoming the nonidentifiability dilemma. *Risk Analysis*, 17(3):367–374, 1997.

Bernard W. Stewart and Paul Kleihues, editors. *World Cancer Report*. IARC *Press*, International Agency for Research on Cancer, World Health Organization, Lyon, 2003.

G. M. Tallis and P. Chesson. Identifiability of mixtures. *J. Austral. Math. Soc. Ser. A*, 32(3):339–348, 1982. ISSN 0263-6115.

Wai-Yuan Tan. Some mixed models of carcinogenesis. *Math. Comput. Modelling*, 10(10):765–773, 1988.

Wai-Yuan Tan. *Stochastic Models of Carcinogenesis*, volume 116 of *STATISTICS: Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1991.

Wai-Yuan Tan and Karan P. Singh. A mixed model of carcinogenesis with applications to retinoblastoma. *Mathematical Biosciences*, 98: 211–225, 1990.

Henry Teicher. Identifiability of finite mixtures. *Annals of Mathematical Statistics*, 34:1265–1269, 1963.

Ian Tomlinson, Peter Sasieni, and Walter Bodmer. How many mutations in a cancer. *American Journal of Pathology*, 160(3):755–758, March 2002.

James W. Vaupel and Anatoli I. Yashin. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *Amer. Statist.*, 39(3):176–185, 1985. ISSN 0003-1305.

Bert Vogelstein and Kenneth W. Kinzler. Cancer genes and the pathways they control. *Nature Medicine*, 10(8):789–799, August 2004.

Robert A. Weinberg. How cancer arises. *Scientific American*, 275: 32–40, September 1996.

Alice Whittemore and Joseph B. Keller. Quantitative theories of carcinogenesis. *SIAM Rev.*, 20(1):1–30, 1978. ISSN 0036-1445.

Jianfeng Xu, James Lowey, Fredrik Wiklund, Jielin Sun, Fredrik Lindmark, Fang-Chi Hsu, Latchezar Dimitrov, Baoli Chang, Aubrey R. Turner, Wennan Liu, Hans-Olov Adami, Edward Suh, Jason H. Moore, S. Lilly Zheng, William B. Isaacs, Jeffrey M. Trent, and Henrik Grönberg. The interaction of four genes in the inflammation pathway significantly predicts prostate cancer risk. *Cancer Epidemiol Biomarkers Prev*, 14(11):2563–2568, 2005.

Qi Zheng. On the exact hazard and survival functions of the mvk stochastic carcinogenesis model. *Risk Analysis*, 14(6):1081–1084, 1994.

Curriculum Vitae

Sandro Gsteiger

Rue Hans-Geiler 1
1700 Fribourg
sandro.gsteiger@epfl.ch

Date of birth: 27.5.1977
Place of birth: Thun
Nationality: Swiss

Education

2002-2006	PhD student at the Chair of Applied Statistics Institute of Mathematics Swiss Federal Institute of Technology Lausanne
1997-2002	Undergraduate studies in mathematics and biology University of Fribourg, Switzerland
1993-1997	Maturity (University entrance): sciences Gymnasium Interlaken, Switzerland

Foreign Stays

2/2005-8/2005	Dipartimento di Matematica, Politecnico di Torino, Italy Fellowship for prospective researchers from the Swiss National Science Foundation
1999-2000	University Montpellier 2, France Undergraduate studies in mathematics (ERASMUS)