Models for the Development of Tumours in Neurofibromatosis 2

by

Ryan R. Woods

B.Sc, University of Guelph 1998

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

we accept this thesis as conforming
to the required standard

The University of British Columbia

June 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date August 2 2002

DE-6 (2/88)

Abstract

Neurofibromatosis 2 (NF2) is a rare genetic disease that affects approximately 1 in 40000 people. some of the characteristic features of this disease include the onset of multiple tumours on the cranial and spinal nerves, juvenile cataracts and hearing loss. Almost all affected individuals develop bilateral tumours of the schwann cells that line the vestibular nerves; these tumours are called as vestibular schwannomas (VS). Evidence from molecular genetic studies has suggested that a "2-hit" hypothesis is appropriate for the development of VS in patients with NF2; that is to say that a tumour cell develops from a normal schwann cell after the cell sustains two mutations to its genetic material. Several authors have proposed probabilistic models for this process and have shown that such models are consistent with incidence data for several different types of cancer.

We will discuss a selection of probabilistic models for a "2-hit" hypothesis and present the results from the fitting of such models to incidence data from NF2 patients. Molecular evidence does not exclude the possibility that additional hits are necessary for the development of VS; we will discuss a "3-hit" model and compare the model's fit to both the data and to the fit of the "2-hit" models. Genotype-phenotype correlations have been reported in patients with NF2 and thus a model that incorporates a patient's genotype is presented and discussed. Finally, a bivariate model is proposed to estimate the distributions of the ages at onset of both the first and second VS.

Contents

List of Tables

List of Figures

Acknowledgements

I would like very much to thank Professor Harry Joe for his much appreciated support and supervision on this, and other projects, over the past two years. I would also like to thank Dr. Jan Friedman for his support, guidance and for providing us with this project. Additionally, I thank the rest of the Friedman lab (in its many forms over the past two years): Jacek, Patricia, Yinshan, Dana, Ravi, and all of the others. I would like to thank Dr. Gareth Evans, Dr. Michael Baser and Dr. Frank Mirz for providing the datasets used in this thesis.

I have had a wonderful two years in this department, owed largely to the Department's faculty and my fellow graduate students; thank you all for making this place what it is. Many thanks to my friend Zhu Rong for the use of his (very helpful) thesis proposal. And of course, a herculian "bao bao" to Lee Shean, my partner in crime, for her support throughout the writing of this.

Shout outs to Bubby, Corky, Pete, Lambo and the rest of the posse in Ontario.

RYAN REGINALD WOODS

The University of British Columbia
July 2000

For Hosh Pesotan

Chapter 1

Introduction

The central objective of this thesis is to outline a selection of probabilistic models for the development of tumour cells and provide an application of such models to data from patients with the disease neurofibromatosis 2 (NF2). In this chapter we will provide some background information about the genetic disease NF2. We will outline some of the terminology and genetic concepts that will be used and discussed throughout this thesis. Section 1.2 will provide an overview of the major contributions to mathematical models for carcinogenesis. A selection of these models will be used in this thesis in an application to data on NF2 patients; Section 1.3 will discuss our specific study objectives and discuss the general features of the models we are interested in applying to our data. Finally, a description of the datasets to be used in our study is given in Section 1.4.

1.1 Neurofibromatosis 2

Neurofibromatosis 2, also called Bilateral Acoustic Neurofibromatosis, is a rare genetic disease that affects approximately 1 in 40000 people. All NF2 patients bear some form of a mutation to the *NF2* gene (italics are conventionally used to denote the name of the gene and Roman type for the name of the disease); this mutation

is present at birth in all NF2 patients. NF2 is a dominant disease. This means that it is caused by one mutant copy of the *NF2* gene. People normally have two copies of every gene (except those on the sex chromosomes in males). People with NF2 have one mutant and one normal copy of the *NF2* gene in every cell of their bodies unless additional mutations have occurred in the process of tumorigenesis, as discussed below. The *NF2* gene is located on chromosome 22q and includes 17 exons. Characteristic features of NF2 include the onset of multiple tumours on the cranial and spinal nerves, juvenile cataracts, headaches and facial weakness. We will refer to the various disease features exhibited by a patient as the patient's phenotype.

Almost all affected individuals develop bilateral tumours of the schwann cells that line the vestibular nerve; such tumours are known as vestibular schwannomas (VS). The presence of one or more VSs can cause loss of balance, hearing loss and a ringing in the ears called tinnitus. These are typically the early symptoms of NF2, which often occur in an individual's teenage years or during their twenties.

Approximately 50% of NF2 cases are new mutations; meaning that the affected person's parents did not have NF2. Additionally, a person with NF2 will have a 50% chance of passing on their mutated *NF2* gene to any of their children. When performing statistical analyses on data from NF2 patients it is important to be able to distinguish the family member who first sought medical attention for their condition from the other family members; this family member is called the proband. In most statistical analyses probands and non-probands will be analyzed separately, particularly if the analyses require the use of information related to the age at onset of various features of the disease. The rationale for this is that once a proband has been brought to medical attention, other members of their family will be examined for features of the disease as well; even if they have not previously shown any of the early symptoms of NF2. These other members of the family may be monitored more closely to detect the onset of various features of NF2. This may include giving these family members MRI scans to detect the onset of tumours before they be-

come symptomatic. Information recorded about the age at which features became apparent in the non-probands will tend to be biased towards earlier ages compared with this same information in the probands. Analyzing probands separately from the other members of the family is an attempt to prevent this potential bias from affecting the results of any analyses.

There are several different general types of mutations of the *NF2* gene that can occur. The three major classes of mutation types which seem to be well represented in our NF2 datasets are: protein truncating mutations; missense mutations; and splice-site mutations. These three different mutation types differ from one another in the effect that they produce on the production of protein. More specific details related to these mutation types will appear in the glossary. We will use the term genotype to refer to the type of mutation of the *NF2* gene borne by an individual. This is somewhat different than the conventional definition of genotype found elsewhere in genetics; genotype more often refers to the entire genetic constitution of an organism. In statistical studies of NF2 patients it is of interest to examine the relationship between both the genotype and phenotype of the patients. Suggestions from both clinical and epidemiological studies, that certain types of mutations of the *NF2* gene produce more severe disease features than others [5, 25], have motivated the study of genotype-phenotype relationships.

Further details about NF2 can be found in the book by Friedman *et al.* [6].

1.2 Models for Carcinogenesis

There have been many contributions to the development of mathematical and probabilistic models for human carcinogenesis. Such models have been used in the past to explain incidence rates for different cancers in populations, as well as to validate hypotheses about the genetic mechanisms that are responsible for tumour development. The common theme that is incorporated into most of these models is that a tumour cell is assumed to be the outcome of a sequence of irreversible events; these

events progressively transform normal tissue cells by some mechanism into tumour cells. A presentation of the main ideas and concepts related to many of these models is provided by Chu [2]; this reference provides a clear non-mathematical formulation of these models. A brief outline of some of these models is provided below.

Perhaps the earliest of these models was the multistage model presented by Armitage and Doll [1]. The Armitage-Doll (AD) model described the transformation process of a normal tissue cell into a tumour cell. The transformation process is represented by a sequence of irreversible changes of state. The change from one state to another represents a cell moving from its present state to a state of further malignancy, until it eventually reaches the final stage; a malignant tumour cell which divides until it becomes a detectable cancer. The AD model has been used to explain the incidence rates of many adult human cancers.

Knudson [12, 13, 8] proposed a two-stage model for cancer initiation to describe the incidence of both sporadic and hereditary retinoblastoma. According to Knudson's model, a tissue cell is transformed into a tumour cell after sustaining two irreversible mutations. This model is often referred to as a "two-hit" model; where the term hit refers to the mutation of an allele in the cell. The first of these mutations is assumed to occur in one of two ways: in hereditary cases of the cancer, individuals inherit the first mutation; in non-hereditary, or sporadic cases, the first mutation occurs by chance. The second mutation is assumed to occur by chance in both groups. This two-mutation model also allows for the cell division of both normal tissue cells and cells that have already sustained a mutation; this is a necessary feature of a realistic model. Previously, the AD model had not taken into account the cellular kinetics of the tissue under study. Knudson's model was shown to be consistent with epidemiological data for retinoblastoma and subsequent genetic analyses have established the validity of this model in retinoblastoma and other forms of human cancers. Further details related to Knudson's model will be discussed in Chapter 2.

Moolgavkar and Venzon [18] introduced an important class of two-mutation models that have been used extensively in epidemiological research; we shall refer to these models as the MV models. These models, like Knudson's, also incorporated the cellular kinetics of the tissue into the model; the MV models however, allow for the division and death of both the normal tissue cells and cells that have already sustained mutations. A subclass of these models is able to incorporate an explicit functional form for the number of tissue cells present in the tissue as a function of age [19, 20, 21, 22]. This is a useful feature of the model as some tissues may grow rapidly during some stages of development and then remain at a fixed size thereafter; such growth patterns can be incorporated into the model by choosing an appropriate function for the the number of tissue cells as a function of age. These models have been shown to be consistent with epidemiological and experimental data for many different types of cancers [22]. The mathematical formulation of the model is also convenient to extend to a three-mutation model, and work has also been done to extend this model to a general number of mutations. Several authors have contributed to generalizing the MV models in various ways; many of these contributions are noted in the references. Details related to both the mathematical formulation and application of some of the MV models will appear in Chapter 3.

In the subsequent section we will discuss our intent to use a selection of these models in an application to data collected on NF2 patients. A description of the data that we intend to use will also follow.

1.3 Study questions to address

The central objective of this thesis is to explore the appropriateness of a two-mutation hypothesis for the development of VSs in patients with NF2. Our motivation for this has come from molecular data that have suggested the appropriateness of such a model [6]. A two-mutation hypothesis for the development of VSs would imply that schwann cells around the vestibular nerve must acquire two mutations in

some fashion in order to develop into tumour cells. Additionally, we are interested in exploring the suitability of a three-mutation hypothesis for the development of VSs. One reason for this is that molecular data do not rule out the possibility that a third mutation may contribute to the development of these tumours. As well, the development of certain cancers have previously been shown to be consistent with both two and three-mutation models. We intend to assess the appropriateness of the aforementioned hypotheses by fitting suitable probabilistic models to patient data; a more thorough discussion of these models appears below.

Figure 1.1 is a depiction of a two-hit model for hereditary tumours; circles in this picture represent cells and the arrows between circles represent possible transitions from one type of cell to another. Greek letters along side the arrows represent the rates of transition between states; these will be described more thoroughly in subsequent sections. A cell that bears a single mutation is capable of division, death, or sustaining a second mutation; cells that have died or sustained the second mutation are not capable of returning to the single mutation state. Under such a model a tumour cell is any tissue cell that has sustained two mutations. This model is applicable to NF2 patients as all NF2 patients are born with a mutation to one *NF2* allele and thus tumours associated with NF2 can be considered hereditary tumours. Figure 1.2 is a picture of a three-hit model for hereditary tumours. This model is identical to the two-hit model described above except that it contains an additional stage; cells that have sustained two mutations are now also capable of division, death, and acquiring a third mutation. A tumour cell is generated when a tissue cell acquires the third mutation. We will provide mathematical formulations for these models in Chapters 2 and 3 and describe the fitting of such models to NF2 data and the results in Chapter 5.

As a starting point, we will provide an outline of Knudson's two-hit model and apply this model to data from both sporadic and hereditary VS, the latter in patients with NF2. Knudson's model however, does not allow the incorporation of

Figure 1.1: 2-hit model for NF2 patients (hereditary tumours)

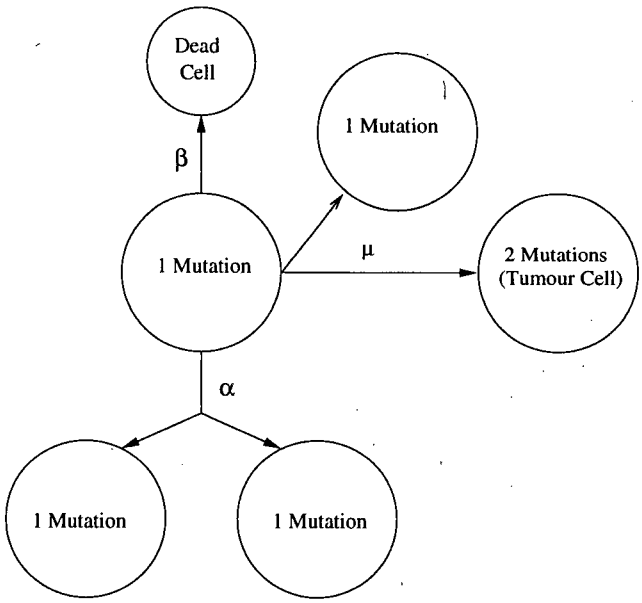
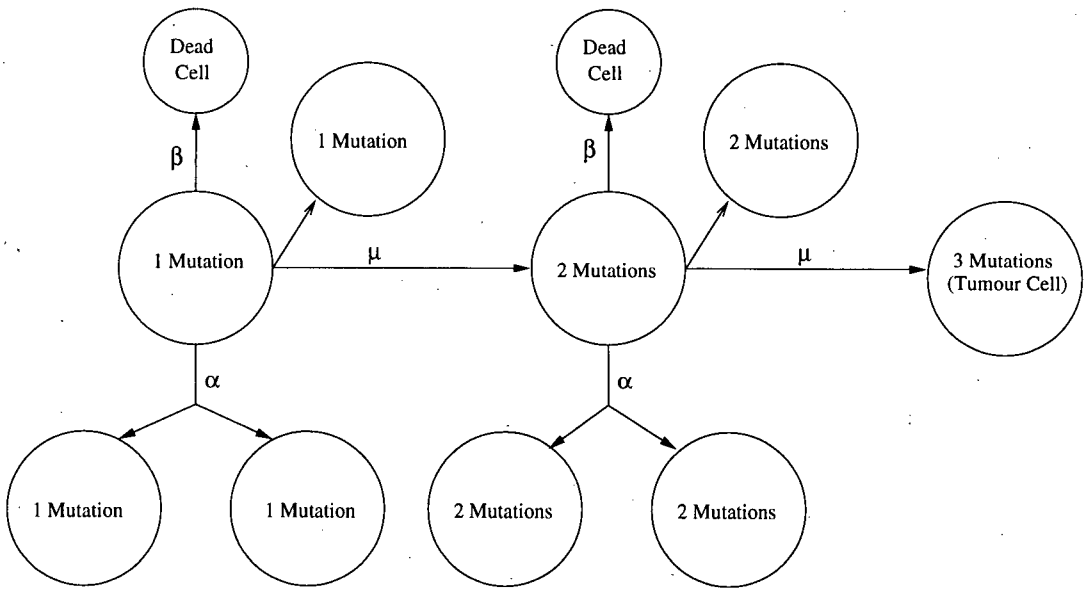


Figure 1.2: 3-hit model for NF2 patients (hereditary tumours)



genotype information into the model. Associations between patients' genotype and phenotype have been described in NF2 [5, 25] and thus we would like to develop a model capable of incorporating genotype information. The MV models described in the previous section appear to be capable of including this information. We will provide an outline of the mathematical formulation of these models and describe our approach to incorporating the genotype effects into these models. These models will be fit to patient data to assess the suitability of the two-mutation hypothesis and to examine the role of genotype in predicting the onset of VS in NF2 patients. We will also provide the mathematical details of a 3-hit model for the development of tumours in NF2 patients. This model will also be fit to data to assess the suitability of a three-mutation hypothesis.

Previously, both Knudson's model and the MV models have been used to predict the incidence of the first tumour in a tissue of interest. Many tissue however are bilateral, or paired, and there may be interest in predicting the age at which a person develops tumours in each of the paired tissues. The vestibular nerve is a bilateral tissue and NF2 patients develop tumours of the schwann cells along both the left and right vestibular nerves. We will present a model that is able to predict the age at which a patient develops both VSs under the assumption that two mutations are necessary for the development of a VS. This model will be presented mathematically in Chapter 4 and the results from fitting this model to data will appear in Chapter 5.

Finally, we will provide some comments about how the fitting of these models could be simplified or improved as well as some other potentially interesting study questions that could be addressed using similar models.

1.4 Description of the data

It is quite difficult to acquire large NF2 data sets as a result of the low prevalence of the disease. This often necessitates combining data from many sources so that

enough patient data are available for statistical analyses. We are fortunate to have access to a reasonably large database of NF2 patient data; the information contained in this database is both clinical and molecular for many of the patients. Unfortunately, for some of our analyses we will still need to combine data from several published sources in order to have sufficient sample sizes for our analyses. Here we will provide a description of the data that will be used to fit our models.

The large database available to us was provided by Dr. Gareth Evans, St. Mary's Hospital, Manchester, U.K.; we will henceforth refer to this database as the MUK (Manchester, U.K.) data. As of March 1, 2000, this database contains detailed clinical information on 349 NF2 patients. Information contained in the patient records includes ages at onset for several characteristic features of NF2, age at presentation of the first feature of NF2, age at last examination, proband status (yes/no), presence or absence of several different types of tumours, laterality of VS, and many other variables. All of the aforementioned information relates to the phenotype of the patient; a subset of the patients also has genotype information recorded. 188 patients from this dataset had their DNA sequenced to determine the type of their germline *NF2* mutation.

To be eligible for our study, patients had to be probands with bilateral VS with the age at onset of the first VS recorded. Using probands for our model fitting is an attempt to remove biases that may result from analyzing data that contain both probands and non-probands; this point was discussed previously in Section 1.1. The NIH diagnostic criteria (1997) for NF2 state that a person can be diagnosed with NF2 only if they satisfy one or more of the following:

- Bilateral VS
- Family history of NF2 and either:

  1) Unilateral VS at age less than 30

  2) Any two of the following: meningioma, glioma, schwannoma, juvenile

posterior subscapsular lenticular opacities/juvenile cortical cataract

Probands by definition will not have a family history of NF2 and thus a proband will only be diagnosed with NF2 following the onset of both VSs. To ensure that the patients we are including in our analyses do in fact meet the NIH crtieria for NF2, we require that they have bilateral VS. The age at onset of the first VS will typically be the response variable for our analyses and thus all patients used in our analyses must have this variable known. In some cases a surrogate measure for this can be used in place of the age at onset of the first tumour and this will be further discussed below. In total, the MUK data contained 163 probands with bilateral VS and a recorded age at onset of first VS variable.

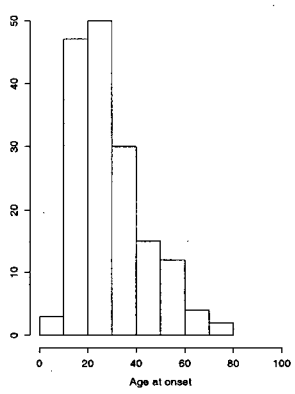
In order to acquire enough patients with known genotype information we combined the MUK data with data from several other published sources. This database has been compiled and maintained by Dr. Mike Baser from Los Angeles, U.S.A.; we will refer to this as the FSS (from several sources) data set. A complication with merging data from these various sources was that not all sources had recorded the age at onset of the first VS for the patients. Instead, several sources recorded the age at onset of hearing loss; this variable has often been used as a surrogate for the age at onset of the first VS as the presence of a VS typically results in a loss of hearing. Fortunately, all data sources had age at onset of hearing loss recorded for their patients and thus any analyses done using the FSS data will use this variable as the dependent variable. To see that age at onset of hearing loss is a suitable surrogate for the age at onset of the first VS one can refer to Figure 1.3. This figure shows both histograms and boxplots for the age at onset of the first VS and its proposed surrogate. Displays are provided for the age at onset of hearing loss from the patients from the MUK data, as well as for patients from the FSS data. The distributions displayed in this figure match one another very closely and suggest the appropriateness of our proposed surrogate measure for the age at onset of the first VS. The FSS data contained 167 patients that met our criteria for entrance

into the study. Of the 167 patients that were eligible for our study, 68 patients from the FSS data had identified mutation types; 40 of these patients had frameshift or nonsense mutations (which we will refer to as protein truncating mutations) and 28 patients had other types of mutations. Figure 1.4 is a display of the distribution of the age at onset of hearing loss by mutation type.

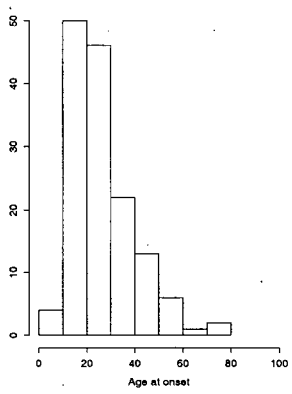
The MUK data had 144 probands with bilateral VS and age at hearing loss recorded. Two patients had not yet developed hearing loss and were omitted from any analyses using age at onset of hearing loss as the dependent variable. These patients could have been regarded as censored in our analyses however, given the small censoring rate we have chosen to remove them to avoid the need to apply methodology for censored data. The influence of two censored observations on our model fitting results would quite surely be negligible given the size of the dataset. It is useful to note that the 144 patients described above are in fact a subset of the FSS dataset.

Finally, we will also employ a dataset of sporadic VS patients. These are patients that do not have NF2, but do have a unilateral VS. The data were provided by Frank Mirz, a researcher from Denmark, and were published in a study on the natural history of VSs [17]. This dataset contains the age at onset of the first VS for 72 patients affected by a unilateral VS. We will henceforth refer to this dataset as the SPOR data. Figure 1.5 shows histograms for the age at onset of the first VS in sporadic and NF2 cases of VS. The distributions are clearly different for the two groups; the onset of the VS occurs much earlier in patients with NF2 than it does in sporadic cases.

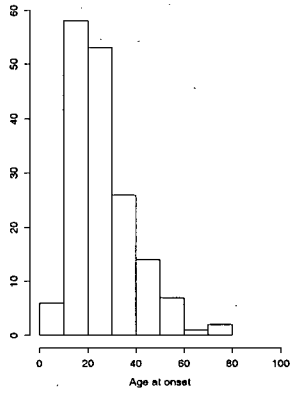
Figure 1.3: Simple descriptive summaries of the three data sets



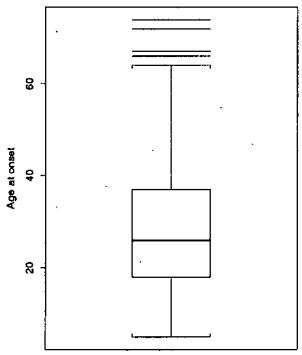
(a) Histogram for the age at onset of the first VS (MUK data)



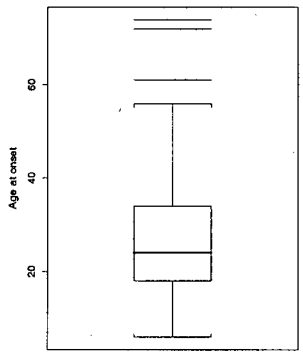
(b) Histogram for the age at onset of hearing loss (MUK data)



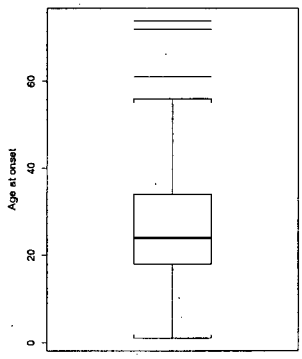
(c) Histogram for the age at onset of hearing loss (FSS data)



(d) Boxplot for the age at onset of the first VS (MUK data)

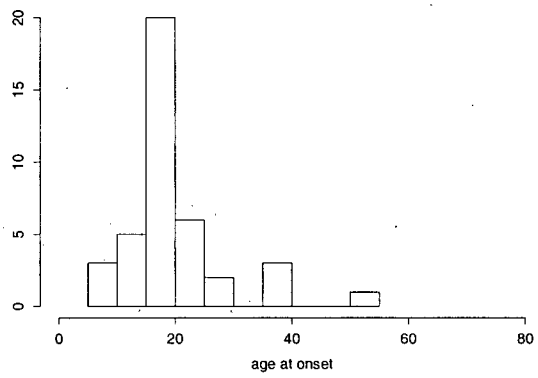


(e) Boxplot for the age at onset of hearing loss (MUK data)

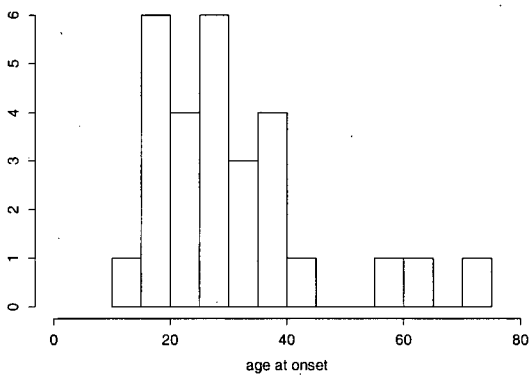


(f) Boxplot for the age at onset of hearing loss (FSS data)

Figure 1.4: Histograms for the age at onset of hearing loss by mutation type (FSS data)

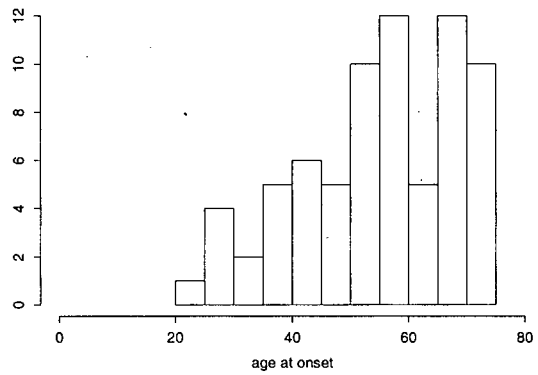


(a) Protein-truncating mutations

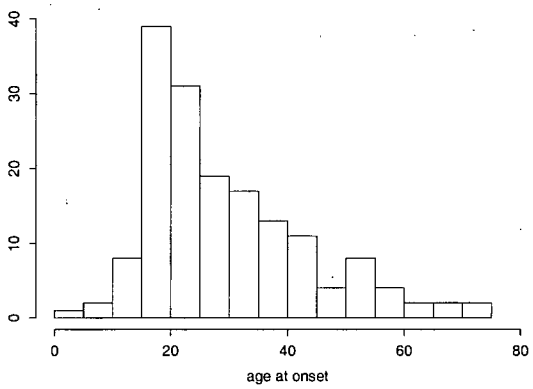


(b) Other identified mutation types

Figure 1.5: Histograms for the age at onset of the first VS



(a) Sporadic cases of VS



(b) NF2 patients (MUK data)

Chapter 2

Knudson's Early Model

Let $H(t)$ and $N(t)$ represent the fraction of undiagnosed cases of hereditary and nonhereditary VS at age t respectively. Recall, that hereditary cases of VS represent NF2 patients and sporadic cases of VS represent patients from the general population who do not have NF2. Hereditary cases will always develop VSs on both the left and right sides of their head; we refer to this condition as bilateral VS. Sporadic cases have only a single VS on one side of their head; it is assumed that either side of the head is equally likely to develop the tumour. We also assume that the two sides of the head develop tumours independently of each other.

We will begin with a mathematical formulation of the model for hereditary cases of VS. Let $m(t)$ be the mean number of tumours, that have developed in the interval $[0, t]$, per individual from the hereditary cases. Our assumption that the two sides of the head are equally likely to develop a tumour suggests that the expected number of tumours that have developed prior to time t, on one side of the head, is $m(t)/2$. We assume no delay between the time when a mutation occurs and the time that the tumour is detected; this justifies the equivalence of the events {patient has k mutations} and {patient has k tumours} for our probability calculations. If we assume that the number of tumours on an individual follows a Poisson distribution

with mean $m(t)$ then it is possible to calculate the following probabilities:

$$\begin{aligned}
\Pr(\text{one side of the head has no mutation in } [0,t]) &= \exp\{\frac{-m(t)}{2}\} \\
\Pr(\text{one side of the head has at least 1 mutation in } [0,t]) &= 1-\exp\{\frac{-m(t)}{2}\} \\
\Pr(\text{patient has no mutation in } [0,t]) &= \exp\{-m(t)\} \\
\Pr(\text{patient has at least 1 mutation in } [0,t]) &= 1-\exp\{-m(t)\}
\end{aligned}$$

Further, let us define $M[t_1,t_2]$ to be the number of mutations that have occurred in the interval $[t_1,t_2]$; note that $M[t_1,t_2]=m(t_2)-m(t_1)$. An expression for $H(t)$ can now be found using the aforementioned assumptions and definitions:

$$\begin{aligned}
H(t) &= \Pr(\text{patient has } M[0,t]=0 \mid \text{patient is eventually bilateral}) \\
&= \frac{\Pr(\text{both sides of head have } M[0,t]=0 \text{ and } M[t,\infty]\geq 1)}{\Pr(\text{both sides of head have } M[0,\infty]\geq 1)} \\
&= \frac{\Pr(\text{one side has } M[0,t]=0)^2 \Pr(\text{one side has } M[t,\infty]\geq 1)^2}{\Pr(\text{one side of head has } M[0,\infty]\geq 1)^2} \\
&= \frac{\exp\{-m(t)\}\left[1-\exp\{-\frac{m(\infty)}{2}+\frac{m(t)}{2}\}\right]^2}{\left[1-\exp\{-\frac{m(\infty)}{2}\}\right]^2} \\
&= \frac{\left[\exp\{-\frac{m(t)}{2}\}-\exp\{-\frac{m(\infty)}{2}\}\right]^2}{\left[1-\exp\{-\frac{m(\infty)}{2}\}\right]^2}.
\end{aligned}$$

The formulation of the model for sporadic cases of VS is similar to that of the hereditary cases. We define $q(t)$ to be the mean number of tumours developed per person, prior to time t, for individuals from the sporadic population. Assuming a Poisson distribution for the random number of tumours per individual, as above, we arrive at the following expression for $N(t)$:

$$N(t)=\frac{\exp\{-q(t)\}-\exp\{-q(\infty)\}}{1-\exp\{-q(\infty)\}} \simeq 1-\frac{q(t)}{q(\infty)},$$

where the approximate equality is justified by considering that $q(t)$ is much smaller

than 1; thus, $1 - \exp\{q(t)\}$ can be approximated by its first order Maclaurin approximation $q(t)$.

In the formulation of the expressions for $H(t)$ and $N(t)$ we have allowed them to depend on t by allowing $m(t)$ and $q(t)$ to be age dependent; thus far we have not described how these functions will depend on age. We will now decribe Knudson's approach of relating the functions $m(t)$ and $q(t)$ to the number of cell divisions that have occurred in the tissue prior to age t.

We must first make several assumptions about the tissue cells and their cellular kinetics. First, it is assumed that both normal tissue cells and cells that have already sustained a single mutation are capable of cell division. It is also necessary to assume that the mean number of cell divisions that have occurred prior to age t is equal in the hereditary and sporadic cases; we denote this quantity by $a(t)$. We denote the number of tissue-specific cells present at time t to be $b(t)$; where the initial number of cells present in the tissue is denoted by $b(0)$. Clearly, the number of tissue cells present at any time t would be equal to the number initially present in the tissue plus the number of cell divisions that have occurred prior to this time: $b(t) = b(0) + a(t)$. Letting $t \to \infty$ we obtain the number of tissue cells that will eventually be present in the tissue: $b(\infty) = b(0) + a(\infty)$. We are now able to define a function $d(t)$, which will represent the fraction of cell divisions that have occurred by time t, in the following fashion:

$$d(t) = \frac{a(t)}{a(\infty)} = \frac{b(t) - b(0)}{b(\infty) - b(0)}.$$

In hereditary cases, the functions $m(t)$ and $d(t)$ are related by assuming that tumour cells arise by transformation of an intermediate cell at a constant mutation rate denoted by μ_2; the units for this rate are given as mutations per cell division. Thus we can express the mean number of tumours developed prior to time t as: $m(t) = \mu_2 \cdot a(t)$. The assumption that the mutation rate is constant with respect to

time leads immediately to the following result:

$$\begin{aligned} \mu_2 &= \frac{m(t)}{a(t)} = \frac{m(\infty)}{a(\infty)} \\ &\Leftrightarrow \frac{m(t)}{m(\infty)} = \frac{a(t)}{a(\infty)} = d(t) \\ &\Leftrightarrow m(t) = d(t)m(\infty). \end{aligned}$$

This relationship can now be substituted into our expression for $H(t)$ given previously to yield:

$$H(t) = \frac{\left[-\exp\left\{\frac{m(\infty)d(t)}{2}\right\} - \exp\left\{-\frac{m(\infty)}{2}\right\}\right]^2}{\left[1 - \exp\left\{-\frac{m(\infty)}{2}\right\}\right]^2}. \tag{2.1}$$

Recall that for a tumour to develop in an individual from the general population, two chance mutations are required: the first to transform a normal tissue cell into an intermediate cell, and the second to transform the intermediate cell into a tumour cell. The first and second mutations are assumed to occur at constant rates of μ_1 and μ_2 respectively; again the units for the mutation rates are given in mutations per cell division. We define $p(t)$ to be the mean number of mutations of normal tissue cells that have occurred prior to time t. Clearly $p(t)$ can be expressed as: $p(t) = \mu_1 \cdot a(t)$. The quantity that we would like to relate to the number of cell divisions is in fact $q(t)$, the number of mutations of intermediate cells that yield tumour cells. This quantity would simply be the product of the number of intermediate cell divisions that have occurred by time t and the mutation rate μ_2. If the number of intermediate cells present in the tissue are represented by $I(t)$, then the number of intermediate cell divisions that have occurred by time t would simply be $I(t) - p(t)$. Thus, we yield the following expression:

$$q(t) = \mu_2[I(t) - p(t)] = \mu_2[I(t) - \mu_1 a(t)].$$

We will now outline Knudson's suggestion for a method of estimating the mean number of intermediate cells present at time t. We begin by partitioning the

interval $[0,t]$ into n subintervals of length h:

$$[0,t] = \bigcup_{i=1}^{n} [\tau_i, \tau_i + h].$$

Recalling the definition of $p(t)$, we note that the mean number of mutations that produce intermediate cells from normal tissue cells in the ith subinterval is simply $p(\tau_i + h) - p(\tau_i)$. These intermediate cells will also divide and increase in number; it is assumed that the proportional increase in the number of intermediate cells from age τ_i to age t is equal to this proportional increase in the normal tissue cells. The proportional increase in the number of intermediate cells from age τ_i to age t is given by:

$$\frac{b(t)}{b(\tau_i)} = \frac{a(t) + b(0)}{a(\tau_i) + b(0)}.$$

A final expression for $I(t)$ can now be obtained by computing the following integral:

$$\begin{aligned}
I(t) &= \lim_{h \to 0} \sum_{i=1}^{n} \frac{p(\tau_i + h) - p(\tau_i)}{h} \frac{b(t)}{b(\tau_i)} h \\
&= \int_0^t p'(\tau) \frac{b(t)}{b(\tau)} d\tau \\
&= \mu_1 \int_0^t a'(\tau) \frac{a(t) + b(0)}{a(\tau) + b(0)} d\tau \\
&= \mu_1 [a(t) + b(0)] \log\{a(\tau) + b(0)\} \Big|_0^t \\
&= \mu_1 [a(t) + b(0)] \{\log\{a(t) + b(0)\} - \log\{b(0)\}\}.
\end{aligned}$$

Note that we have used the fact that $a(0) = 0$ in the final equality above; a reasonable assumption of course as $a(0)$ represents the number of cell divisions that have occurred prior to age 0. Substituting this expression into our expression for $q(t)$, and using the fact that for positive t, $a(t)$ will be much larger than $b(0)$, we yield:

$$\begin{aligned}
q(t) &= \mu_2 \Big\{ \mu_1 [a(t) + b(0)] \{\log\{a(t) + b(0)\} - \log\{b(0)\}\} - \mu_1 a(t) \Big\} \\
&\simeq \mu_1 \mu_2 a(t) \Big[\log\Big\{ \frac{a(t)}{b(0)} \Big\} - 1 \Big].
\end{aligned}$$

Substituting this expression into our expression for $N(t)$, and recalling that $d(t) = a(t)/a(\infty)$, leads to:

$$\begin{aligned}
N(t) &\simeq 1 - \frac{q(t)}{q(\infty)} \\
&\simeq 1 - \frac{\mu_1\mu_2 a(t)\left[\log\left\{\frac{a(t)}{b(0)}\right\} - 1\right]}{\mu_1\mu_2 a(\infty)\left[\log\left\{\frac{a(\infty)}{b(0)}\right\} - 1\right]} \\
&= 1 - d(t)\frac{\left[\log\left\{\frac{a(t)}{b(0)}\right\} - 1\right]}{\left[\log\left\{\frac{a(\infty)}{b(0)}\right\} - 1\right]} \\
&= 1 - d(t)\frac{\left[\log\left\{\frac{a(t)a(\infty)}{b(0)a(\infty)}\right\} - 1\right]}{\left[\log\left\{\frac{a(\infty)}{b(0)}\right\} - 1\right]} \\
&= 1 - d(t)\frac{\left[\log\left\{\frac{a(t)}{a(\infty)}\right\} + \log\left\{\frac{a(\infty)}{b(0)}\right\} - 1\right]}{\left[\log\left\{\frac{a(\infty)}{b(0)}\right\} - 1\right]} \\
&= 1 - d(t)\frac{\log\{d(t)\}}{\left[\log\left\{\frac{a(\infty)}{b(0)}\right\} - 1\right]} - d(t). \qquad (2.2)
\end{aligned}$$

The result of the previous mathematical formulation is that both functions $H(t)$ and $N(t)$ are made to depend on t only through the function $d(t)$. This function is a time-dependent parameter which can be estimated from the data at various time points. $H(t)$ and $N(t)$ also depend on two other quantities as well; values of $m(\infty)$ and $a(\infty)/b(0)$ can be selected prior to the analysis leaving the time-dependent parameter $d(t)$ as the only unknown parameter to estimate from the data.

Suppose we want to estimate the fraction of cell divisions that have occurred prior to k different times $t_1, ..., t_k$. Suppose, additionally that we have n_h and n_s hereditary and sporadic cases respectively. We denote the observed ages at onset of VS in the hereditary group as $t_1^h, ..., t_{n_h}^h$; these same ages in the sporadic group are denoted by $t_1^s, ..., t_{n_s}^s$. To obtain an estimate of $d(t_i)$ $(i = 1, ..., k)$, Hethcote *et al.*

suggested the minimization of the following function Q with respect to $d(t_i)$:

$$\begin{aligned} Q &= W_h(t_i)\Big\{\widehat{H(t_i)} - \frac{\left[\exp\{-\frac{m(\infty)d(t_i)}{2}\} - \exp\{-\frac{m(\infty)}{2}\}\right]^2}{\left[1-\exp\{-\frac{m(\infty)}{2}\}\right]^2}\Big\}^2 \\ &+ W_s(t_i)\Big\{\widehat{N(t_i)} - 1 + d(t_i) + \frac{d(t_i)\log\{d(t_i)\}}{\log\{\frac{a(\infty)}{b(0)}\} - 1}\Big\}^2, \end{aligned}$$

where $\widehat{H(t_i)}$ and $\widehat{N(t_i)}$ are estimated by the empirical survival functions:

$$\begin{aligned} \widehat{H(t_i)} &= \frac{\#\{t_j^h > t_i\}}{n_h}, \quad j = 1, ..., n_h, \\ \widehat{N(t_i)} &= \frac{\#\{t_j^s > t_i\}}{n_s}, \quad j = 1, ..., n_s, \end{aligned}$$

and $W_h(t_i)$ and $W_s(t_i)$ are weights used in the minimization; the values chosen for the weights are the number of undiagnosed hereditary and sporadic cases at time t_i repectively. Thus,

$$\begin{aligned} W_h(t_i) &= \#\{t_j^h > t_i\}, \quad j = 1, ..., n_h, \\ W_s(t_i) &= \#\{t_j^s > t_i\}, \quad j = 1, ..., n_s. \end{aligned}$$

The minimization described above can be quite easily be performed numerically. The fit of the model can be assessed both by a chi-square goodness of fit test and also by examining plots of empirical and model predicted incidence curves for their agreement. These topics will be addressed further in Chapter 5.

Chapter 3

Multi-hit Models and a Maximum Likelihood Approach

For the maximum likelihood approach to multi-hit models we derive the hazard function (also known as the hazard rate function) for the time to the generation of the first tumour cell. We assume that our tumour onset times follow this distribution and construct a likelihood in the customary fashion. There are two two-hit models presented below; the first model has a single parameter and a closed form maximum likelihood estimate can be obtained. The second two-hit model has several parameters and a more complicated expression for the hazard and thus the maximum likelihood estimation is carried out numerically; we provide the hazard function necessary to construct the likelihood. Similarly, for the three mutation model we derive the hazard function for the time to the first tumour and estimates of the model parameters are found numerically. A general approach to constructing the likelihood from a hazard function is also provided. The models presented in this chapter are models for the development of tumours in patients with NF2; this is an important point as it implies that cells in the tissue at risk have already sustained a single mutation.

3.1 Notation

Notation used throughout the remaining sections has been chosen to be consistent with the major references cited. Notation for variables, functions and parameters will be defined as these concepts are introduced. Figures 1.1 and 1.2 were depictions of the processes that we aim to model in this chapter and some simple notation was defined previously with respect to these figures; we will retain this notation throughout the thesis and we redefine it here for convenience. Some simple guidelines for notation are given below:

- $\boldsymbol{\theta}$ will denote the parameter vector for the model under discussion. In most cases this vector will contain the growth, death, and mutation rates for the 2-hit or 3-hit model, e.g. $\boldsymbol{\theta} = (\alpha, \beta, \mu)$.
- α will denote a growth rate or cell division rate for the tissue cells.
- β will denote a death rate for the tissue cells.
- μ will denote a mutation rate for the tissue cells. This represents the rate at which cells with 1 mutation transform into cells with 2 mutations (or the rate by which cells with 2 mutations transform into cells with 3 mutations).
- $h(t|\boldsymbol{\theta})$ will denote the hazard function for the random variable T (representing the time to the first tumour):

$$h(t|\boldsymbol{\theta}) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t, \boldsymbol{\theta})}{\Delta t}.$$

  Additionally, $f(t|\boldsymbol{\theta}), S(t|\boldsymbol{\theta})$, and $F(t|\boldsymbol{\theta})$ will be used to denote the density, survival, and distribution functions for T respectively.

- t_i will denote the realization of the random variable T for patient i. In our applications this will be the age at onset of the first VS or in some cases the age at onset of hearing loss.

In general all of the rate parameters could be subscripted to denote the stage of the model to which they belong (example: μ_1 would represent the rate at which cells transform from normal tissue cells into cells with a single mutation; μ_2 would represent the rate at which cells with a single mutation transform into cells with two mutations, etc.). In our applications however, we will assume that rates of the same kind from different stages of the model are equal (i.e. $\mu_1 = \mu_2 = \mu$).

3.2 2-hit Models

There are two approaches to 2-hit models presented in subsequent sections. The two approaches differ in the assumptions about the growth and death of the tissue cells. The first model assumes that the tissue cells grow according to a deterministic process, and thus the number of cells present in the tissue at any given time t is given by a function $X(t)$; additionally, it is assumed that there is a small chance that any of the cells mutate. The second model assumes that tissue cells divide, die and mutate according to a birth-death process and thus the number of tissue cells present at any time t is a random variable $X(t)$. I have used the same notation for both and have let the context of the discussion distinguish between the random variable $X(t)$ and the deterministic function $X(t)$.

3.2.1 Deterministic growth of tissue

If the tissue cells are assumed to grow according to a deterministic process then the generation of mutated cells can be modelled as a nonhomogeneous Poisson process.

Let $X(s)$ represent the number of tissue cells at time s; note that these tissue cells have already sustained a single mutation in patients with NF2. Further, let $Z(s)$ represent the number of tumour cells in the tissue at time s. These cells have sustained 2 mutations; one which has been inherited and the other which has occurred by chance. The mutation rate of the second chance mutation is assumed to be constant and is denoted by μ. The mean number of tumour cells that have

developed by time t, $\mathrm{E}(Z(t))$, is simply $\int_0^t \mu X(s)ds$. The number of tumour cells can be thought of as arising from a nonhomogeneous Poisson process with intensity $\mu X(t)$.

Let T be the time to the generation of the first tumour cell. The time to the generation of the first tumour cell can be shown to have the following distribution function:

$$\begin{aligned} F(t|\boldsymbol{\theta}) &= \Pr(T \leq t) = 1 - \Pr(T > t) = 1 - \Pr(Z(t) = 0) \\ &= 1 - \exp\Big\{-\mu \int_0^t X(s)ds\Big\}, \quad t \geq 0. \end{aligned}$$

The probability density function for the random variable T is obtained by differentiating the distribution function:

$$f(t|\boldsymbol{\theta}) = \mu X(t) \exp\Big\{-\mu \int_0^t X(s)ds\Big\}, \quad t \geq 0.$$

Hence the hazard function for the time to the first tumour cell, $h(t|\boldsymbol{\theta})$, is simply:

$$h(t|\boldsymbol{\theta}) = \frac{f(t|\boldsymbol{\theta})}{S(t|\boldsymbol{\theta})} = \frac{\mu X(t) \exp\Big\{-\mu \int_0^t X(s)ds\Big\}}{\exp\Big\{-\mu \int_0^t X(s)ds\Big\}} = \mu X(t), \quad t \geq 0.$$

The parameter vector $\boldsymbol{\theta}$ for this model contains only a single parameter; namely the mutation rate parameter μ as the growth and death of cells in the tissue are accounted for in the deterministic growth function for the tissue. A likelihood for the data can be constructed based on the distribution of the time to the onset of the tumour and a maximum likelihood estimate for the rate parameter μ can easily be obtained. If the data consist of n individuals with onset times $t_1, ..., t_n$, the likelihood, $L(\boldsymbol{\theta})$, is given by:

$$L(\boldsymbol{\theta}) = \mu^n \left(\prod_{i=1}^n X(t_i)\right) \exp\Big\{-\mu \sum_{i=1}^n \left(\int_0^{t_i} X(s)ds\right)\Big\}. \tag{3.1}$$

The maximum likelihood estimate of the rate parameter μ has a simple closed form solution given by:

$$\hat{\mu} = \frac{n}{\sum_{i=1}^n \left(\int_0^{t_i} X(s)ds\right)} \tag{3.2}$$

After having estimated this rate parameter it is possible to compute an estimate of the probability that an individual will develop a tumour before a given time t; this is done by replacing μ by its estimate in the expression for the distribution function $F(t|\boldsymbol{\theta})$ given above in equation (3.1). These estimates can be compared to an empirical distribution function for the data to assess the fit of the model.

An asymptotic standard error for $\hat{\mu}$, based on the observed Fisher information, can be estimated by twice differentiating the log-likelihood $l(\mu)$:

$$\widehat{SE} = \frac{1}{\sqrt{-l''(\hat{\mu})}} = \sqrt{\frac{\hat{\mu}^2}{n}}.$$

3.2.2 Stochastic growth of tissue

For this model, we assume that the tissue cells divide, die and mutate according to a birth-death process. We derive the hazard function for the time to first tumour starting from a single tissue cell: $h(t|\boldsymbol{\theta})$. The hazard function for the time to first tumour starting from a tissue consisting of N cells would just be $N \cdot h(t|\boldsymbol{\theta})$ because of the assumption that the cells mutate independently of one another. Additional model assumptions and a derivation of the hazard function $h(t|\boldsymbol{\theta})$ are provided below.

The growth, death and mutation of the tissue cells are assumed to follow a process similar to a birth-death process. In a small interval of time Δt, a tissue cell may divide into two tissue cells with probability $\alpha\Delta t + o(\Delta t)$; die with probability $\beta\Delta t + o(\Delta t)$; divide into a normal tissue cell and a tumour cell with probability $\mu\Delta t + o(\Delta t)$; the probability of more than one such event occurring in this time interval is $o(\Delta t)$. Additionally, we assume that a tumour arises from a single progenitor tumour cell and that the tissue cells mutate independently of one another. Mutations are assumed to occur during cell division; such a cell division will produce both a cell identical to the original progenitor and a cell that has sustained an additional mutation. This assumption is used in the derivation of the Kolmogorov forward equation below.

As previously, we define $X(t)$ and $Z(t)$ to be the number of tissue and tumour cells present in the tissue at time t respectively. Let $\phi(x,z;t)$ be the probability generating function for the number of tissue and tumour cells at time t starting with a single tissue cell initially. Thus,

$$\phi(x,z;t) = \sum_{j,k=0}^{\infty} p((j,k);t)x^j z^k$$

where

$$p((j,k);t) = \Pr(X(t) = j, Z(t) = k | X(0) = 1, Z(0) = 0).$$

Note that we assume:

$$p((j,k);t) = 0 \text{ for } j < 0 \text{ or } k < 0.$$

Our claim is that the hazard function for the time to the generation of the first tumour cell T, is given by the expression $h(t|\boldsymbol{\theta}) = -\phi'(1,0;t)/\phi(1,0;t)$. To see this is true we make the following calculations:

$$\phi(1,0;t) = \sum_{j=0}^{\infty} \Pr(X(t) = j, Z(t) = 0) = \Pr(Z(t) = 0), \tag{3.3}$$

and

$$\begin{aligned}
\phi'(1,0;t) &= \frac{\partial \phi(x,z;t)}{\partial t}\Big|_{x=1,z=0} \\
&= \sum_{j=0}^{\infty} \frac{\partial \Pr(X(t) = j, Z(t) = 0)}{\partial t} = \frac{\partial \Pr(Z(t) = 0)}{\partial t} \\
&= \lim_{\Delta t \to 0} \frac{\Pr(Z(t+\Delta t) = 0) - \Pr(Z(t) = 0)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Pr(Z(t+\Delta t) = 0, Z(t) = 0) - \Pr(Z(t) = 0)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Big\{\Pr(Z(t+\Delta t) = 0 \mid Z(t) = 0) - 1\Big\} \Pr(Z(t) = 0)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{\Big\{-\Pr(Z(t+\Delta t) = 1 \mid Z(t) = 0)\Big\} \Pr(Z(t) = 0)}{\Delta t} \\
&= \Pr(Z(t) = 0) \lim_{\Delta t \to 0} \frac{-\Pr(Z(t+\Delta t) = 1 \mid Z(t) = 0)}{\Delta t} \\
&= -\phi(1,0;t) \lim_{\Delta t \to 0} \frac{\Pr(Z(t+\Delta t) = 1 \mid Z(t) = 0)}{\Delta t},
\end{aligned} \tag{3.4}$$

where this last line is justified using equation (3.3). Recall that T denotes the random time until the generation of the first tumour cell. Rearranging the result from equation (3.4) gives us the following result:

$$\begin{aligned} -\frac{\phi'(1,0;t)}{\phi(1,0;t)} &= \lim_{\Delta t \to 0} \frac{\Pr(Z(t+\Delta t)=1 \mid Z(t)=0)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t+\Delta t \mid T \geq t)}{\Delta t} \\ &= h(t|\boldsymbol{\theta}), \end{aligned}$$

which is the hazard function for the random variable T.

We will now derive a closed form solution for the hazard function for this model. As we have just shown, this will require a solution for the probability generating function $\phi(x,z;t)$ defined above. We begin by defining:

$$p_{(j,k)}((m,n);\Delta t) = \Pr(X(t+\Delta t)=m, Z(t+\Delta t)=n | X(t)=j, Z(t)=k).$$

From the Chapman-Kolmogorov equations we yield:

$$\begin{aligned} p((j,k);t+\Delta t) &= p_{(j,k)}((j,k);\Delta t)\cdot p((j,k);t) \\ &\quad + p_{(j-1,k)}((j,k);\Delta t)\cdot p((j-1,k);t) \\ &\quad + p_{(j+1,k)}((j,k);\Delta t)\cdot p((j+1,k);t) \\ &\quad + p_{(j,k-1)}((j,k);\Delta t)\cdot p((j,k-1);t). \end{aligned}$$

We note that:

$$\begin{aligned} p_{(j-1,k)}((j,k);\Delta t) &= (j-1)\alpha\Delta t + o(\Delta t), \\ p_{(j+1,k)}((j,k);\Delta t) &= (j+1)\beta\Delta t + o(\Delta t), \\ p_{(j,k-1)}((j,k);\Delta t) &= j\mu\Delta t + o(\Delta t), \\ p_{(j,k)}((j,k);\Delta t) &= (1 - j\alpha\Delta t - j\beta\Delta t - j\mu\Delta t - o(\Delta t)). \end{aligned}$$

The Kolmogorov forward differential equation can be obtained as:

$$\begin{aligned} p'((j,k);t) &= \lim_{\Delta t \to 0} \frac{p((j,k);t+\Delta t) - p((j,k);t)}{\Delta t} \\ &= -j(\alpha+\beta+\mu)\cdot p((j,k);t) + (j-1)\alpha \cdot p((j-1,k);t) \\ &\quad +(j+1)\beta \cdot p((j+1,k);t) + j\mu \cdot p((j,k-1);t). \end{aligned}$$

We use this result to yield the following differential equation:

$$\begin{aligned} \phi'(x,z;t) &= \frac{\partial \phi(x,z;t)}{\partial t} \\ &= \sum_{j,k=0}^{\infty} p'((j,k);t) x^j z^k \\ &= \sum_{j,k=0}^{\infty} \Big[-j(\alpha+\beta+\mu)\cdot p((j,k);t) \\ &\quad +(j-1)\alpha \cdot p((j-1,k);t) + (j+1)\beta \cdot p((j+1,k);t) \\ &\quad +j\mu \cdot p((j,k-1);t) \Big] x^j z^k \\ &= \Big[\mu x z + \alpha x^2 + \beta - (\alpha+\beta+\mu)x \Big] \frac{\partial \phi(x,z;t)}{\partial x} \end{aligned} \tag{3.5}$$

with initial condition $\phi(x,z;0) = x$. It is obvious that:

$$\phi'(1,0;t) = \Big[\alpha+\beta-(\alpha+\beta+\mu) \Big] \frac{\partial \phi(1,0;t)}{\partial x} = -\mu \cdot \frac{\partial \phi(1,0;t)}{\partial x}.$$

The previous differential equations would be useful for finding a general case solution for $\phi(x,z;t)$ and the computation of moments, however we are interested only in obtaining an expression for the hazard function $h(t|\boldsymbol{\theta})$. Thus, we are interested in computing only $\phi(1,0;t)$ and $\phi'(1,0;t)$. We have included these equations nevertheless for completeness. Finding an expression for the hazard function is simplified by using an observation of Moolgavkar *et al.* [18]; specifically that $\phi(x,z;t)$ can be shown to satisfy the following Riccati differential equation:

$$\phi'(x,z;t) = \alpha\phi^2(x,z;t) + [\mu z - (\alpha+\beta+\mu)]\phi(x,z;t) + \beta. \tag{3.6}$$

This equation can be obtained by considering the integral equation solution for

$\phi(x,z;t)$ given by Moolgavkar *et al.* [18]:

$$\begin{aligned}\phi(x,z;t) &= z\exp\{-\kappa t\} \\ &\quad + \int_0^t \Big[\alpha\phi^2(x,z;t-u) + \beta + \mu z\phi(x,z;t-u)\Big]\exp\{-\kappa u\}du \\ &= z\exp\{-\kappa t\} \\ &\quad + \int_0^t \Big[\alpha\phi^2(x,z;u) + \beta + \mu z\phi(x,z;u)\Big]\exp\{-\kappa(t-u)\}du\end{aligned}$$

where $\kappa = (\alpha+\beta+\mu)$. Multiplying both sides of this equation by $\exp\{\kappa t\}$ and differentiating with respect to t yields:

$$\begin{aligned}&\phi'(x,z;t)\exp\{\kappa t\} + \phi(x,z;t)\kappa\exp\{\kappa t\} \\ =\ & \frac{d}{dt}\Big\{\int_0^t \Big[\alpha\phi^2(x,z;u) + \mu z\phi(x,z;u)\Big]\exp\{\kappa u\}du\Big\} \\ &+\frac{d}{dt}\Big\{\int_0^t \beta\exp\{\kappa(t-u)\}du\Big\} \\ =\ & \frac{d}{dt}\Big\{\int_0^t \Big[\alpha\phi^2(x,z;u) + \mu z\phi(x,z;u)\Big]\exp\{\kappa u\}du\Big\} \\ &+\frac{d}{dt}\Big\{\frac{-\beta}{\kappa}[1-\exp\{\kappa t\}]\Big\} \\ =\ & \Big[\alpha\phi^2(x,z;t) + \mu z\phi(x,z;t)\Big]\exp\{\kappa t\} + \beta\exp\{\kappa t\}. \qquad (3.7)\end{aligned}$$

Rearranging equation (3.7) and multiplying both sides of the equation by $\exp\{-\kappa t\}$ yields the following:

$$\begin{aligned}&\phi'(x,z;t)\exp\{\kappa t\} = -\phi(x,z;t)\kappa\exp\{\kappa t\} + \beta\exp\{\kappa t\} \\ &+\Big[\alpha\phi^2(x,z;t) + \mu z\phi(x,z;t)\Big]\exp\{\kappa t\} \\ \Leftrightarrow\ & \phi'(x,z;t) = -\phi(x,z;t)\kappa + \beta + \Big[\alpha\phi^2(x,z;t) + \mu z\phi(x,z;t)\Big] \\ \Leftrightarrow\ & \phi'(x,z;t) = \alpha\phi^2(x,z;t) + \Big[\mu z - (\alpha+\beta+\mu)\Big]\phi(x,z;t) + \beta\end{aligned}$$

which is the Riccati equation given in equation (3.6) (more details can be found in the appendix). Thus, evaluating this equation at $x=1$ and $z=0$ yields:

$$\phi'(1,0;t) = \alpha\phi^2(1,0;t) - (\alpha+\beta+\mu)\phi(1,0;t) + \beta.$$

Rearranging this equation yields:

$$\frac{\partial\phi(1,0;t)}{\alpha\phi^2(1,0;t)-(\alpha+\beta+\mu)\phi(1,0;t)+\beta}=\partial t$$

$$\Leftrightarrow \qquad \frac{\partial\phi(1,0;t)}{(\phi(1,0;t)-C_1)(\phi(1,0;t)-C_2)}=\alpha\partial t, \tag{3.8}$$

where $C_1 < C_2$ are distinct roots of the polynomial in q:

$$q^2-\frac{(\alpha+\beta+\mu)}{\alpha}q+\frac{\beta}{\alpha}.$$

The roots of this polynomial are easily obtained by applying the quadratic formula:

$$\begin{aligned} C_1 &= \frac{1}{2\alpha}(\alpha+\beta+\mu)-\frac{1}{2\alpha}\sqrt{\alpha^2-2\alpha\beta+2\alpha\mu+\beta^2+2\beta\mu+\mu^2}, \\ C_2 &= \frac{1}{2\alpha}(\alpha+\beta+\mu)+\frac{1}{2\alpha}\sqrt{\alpha^2-2\alpha\beta+2\alpha\mu+\beta^2+2\beta\mu+\mu^2}. \end{aligned} \tag{3.9}$$

The differential equation given above in equation (3.8) can be integrated directly using partial fractions:

$$\int\frac{\partial\phi(1,0;t)}{(\phi(1,0;t)-C_1)(\phi(1,0;t)-C_2)}=\int\alpha\partial t$$

$$\Leftrightarrow \quad \frac{1}{C_1-C_2}\Big\{\int\frac{\partial\phi(1,0;t)}{\phi(1,0;t)-C_1}-\int\frac{\partial\phi(1,0;t)}{\phi(1,0;t)-C_2}\Big\}=\int\alpha\partial t$$

$$\Leftrightarrow \quad \log\big\{\phi(1,0;t)-C_1\big\}-\log\big\{\phi(1,0;t)-C_2\big\}=\alpha(C_1-C_2)t+C'$$

$$\Leftrightarrow \quad \frac{\phi(1,0;t)-C_1}{\phi(1,0;t)-C_2}=C\exp\big\{\alpha(C_1-C_2)t\big\}$$

$$\Leftrightarrow \quad \phi(1,0;t)=\frac{C_1-C_2\left[C\exp\big\{\alpha(C_1-C_2)t\big\}\right]}{1-\left[C\exp\big\{\alpha(C_1-C_2)t\big\}\right]},$$

where C is a constant to be determined by the initial condition of the differential equation $\phi(1,0;0)=1$. Evaluating our solution for $\phi(1,0;t)$ above at $t=0$ leads us to:

$$\phi(1,0;0)=1 \quad \Leftrightarrow \quad \frac{C_1-C_2C}{1-C}=1 \quad \Leftrightarrow \quad C=\frac{1-C_1}{1-C_2}.$$

Thus our final solution for $\phi(1,0;t)$ is given by:

$$\phi(1,0,t) = \frac{C_1 - C_2\left[\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)t\}\right]}{1-\left[\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)t\}\right]}. \tag{3.10}$$

We can now substitute this into the Riccati equation given previously to obtain our final solution for the hazard function:

$$\begin{aligned} h(t|\boldsymbol{\theta}) &= -\frac{\phi'(1,0;t)}{\phi(1,0;t)} = \frac{-\alpha\phi^2(1,0;t) + (\alpha+\beta+\mu)\phi(1,0;t) - \beta}{\phi(1,0,t)} \\ &= -\alpha\phi(1,0,t) + (\alpha+\beta+\mu) - \beta(\phi(1,0,t))^{-1}. \end{aligned}$$

3.3 3-hit Model

For this model, the tissue is assumed to grow deterministically. It is important to remember that tissue cells already bear a single mutation. Tissue cells mutate according to a random process to produce intermediate cells; an intermediate cell is a cell that has sustained two mutations. The growth, death and mutation of these intermediate cells are assumed to follow a birth-death process. Mutation of an intermediate cell results in a tumour cell; a tumour cell in this model is a cell that has sustained three mutations.

A fully stochastic model could also be proposed here where the tissue cells are assumed to divide, mutate and die according to a birth-death process as well. The problem with such a model is that the mathematics become quite challenging and there is no solution in the literature for the hazard function from such a model. Several authors have proposed an approximate solution to a hazard function from a fully stochastic model; the assumptions necessary for the approximation however, are inappropriate for our application of the model to NF2 patients. The reason for this is that the approximation requires that the probability of tumour in the population of study be very small; in the study of many types of cancer in the general population such an assumption may be quite reasonable. The population

of NF2 patients however, have a very high probability of tumour and thus the application of the approximate hazard function would not be appropriate in the analysis of NF2 data.

Again, we denote the number of tissue cells present at time s as $X(s)$; note that $X(s)$ is a deterministic function. In a small interval of time Δt, the probability that an intermediate cell is generated by the mutation of a tissue cell is $\mu X(s)\Delta t + o(\Delta t)$. The probability that more than one intermediate cell is generated in this fashion is $o(\Delta t)$. The growth, death and mutation of the intermediate cells are assumed to follow a process similar to a birth-death process. Again, in a small interval of time Δt, an intermediate cell divides into two intermediate cells with probability $\alpha\Delta t + o(\Delta t)$; dies with probability $\beta\Delta t + o(\Delta t)$; divides into an intermediate cell and a tumour cell with probability $\mu\Delta t + o(\Delta t)$; the probability of more than one such event occurring in this time interval is $o(\Delta t)$. Additionally, we assume that a tumour arises from a single progenitor tumour cell and that the tissue cells mutate independently of one another. Mutations are assumed to occur during cell division; such a cell division will produce both a cell identical to the original progenitor and a cell that has sustained an additional mutation. This assumption is used in the derivation of the Kolmogorov forward equation below.

We define $Y(t)$ and $Z(t)$ to be the number of intermediate and tumour cells present in the tissue at time t respectively. Let $\Psi(y, z; t)$ be the probability generating function for the number of intermediate and tumour cells. Thus,

$$\Psi(y, z; t) = \sum_{j,k=0}^{\infty} p((j,k); t) y^j z^k,$$

where

$$p((j,k); t) = \Pr(Y(t) = j, Z(t) = k \mid Y(0) = 0, Z(0) = 0).$$

Note that it is assumed that:

$$p((j,k); t) = 0 \text{ for } j < 0 \text{ or } k < 0.$$

We will now derive a closed form solution for the hazard function for this model. An argument identical to that provided in Section 3.2.2 can be used to show that the hazard function is given by the expression $-\Psi'(1,0;t)/\Psi(1,0;t)$. Again, it is necessary for us to find the solution for the probability generating function $\Psi(1,0;t)$. We must first define:

$$p_{(j,k)}((m,n);\Delta t) = \Pr(Y(t+\Delta t) = m, Z(t+\Delta t) = n \mid Y(t) = j, Z(t) = k).$$

From the Chapman-Kolmogorov equations we obtain:

$$\begin{aligned} p((j,k);t+\Delta t) &= p_{(j,k)}((j,k);\Delta t)\cdot p((j,k);t) \\ &\quad + p_{(j-1,k)}((j,k);\Delta t)\cdot p((j-1,k);t) \\ &\quad + p_{(j+1,k)}((j,k);\Delta t)\cdot p((j+1,k);t) \\ &\quad + p_{(j,k-1)}((j,k);\Delta t)\cdot p((j,k-1);t). \end{aligned}$$

We note that:

$$\begin{aligned} p_{(j-1,k)}((j,k);\Delta t) &= (\mu X(t)\Delta t + o(\Delta t)) + ((j-1)\alpha\Delta t + o(\Delta t)), \\ p_{(j+1,k)}((j,k);\Delta t) &= (j+1)\beta\Delta t + o(\Delta t), \\ p_{(j,k-1)}((j,k);\Delta t) &= j\mu\Delta t + o(\Delta t), \\ p_{(j,k)}((j,k);\Delta t) &= 1 - j\alpha\Delta t - j\beta\Delta t - j\mu\Delta t \\ &\quad -\mu X(t)\Delta t - o(\Delta t) \end{aligned}$$

The four equations given above arise as a result of the model assumptions. The first equality holds because if at the beginning of the short time interval there are $j-1$ intermediate cells and k tumour cells, and at the end of the time interval there are j intermediate cells and k tumour cells, then a single intermediate cell has been produced in some fashion. This can occur by the mutation of a tissue cell or by the normal division of an existing intermediate cell. Consideration of these two events will lead directly to the stated probability. The second equality is obtained by considering that the reduction in the number of intermediate cells from $j+1$ to

j can occur only if a current intermediate cell dies. The increase in the number of tumour cells from $k-1$ to k occurs with the mutation of a current intermediate cell; this provides the motivation for the third equation. Recall that such a mutation occurs during cell division and results in both an intermediate cell and a tumour cell; thus, the number of intermediate cells does not change with the occurrence of such an event. The final equality is justified by considering the probability that none of the aforementioned events takes place in the time interval.

The Kolmogorov forward differential equation can be obtained as:

$$\begin{aligned} p'((j,k);t) &= \lim_{\Delta t \to 0} \frac{p((j,k);t+\Delta t) - p((j,k);t)}{h} \\ &= -j(\alpha+\beta+\mu)\cdot p((j,k);t) - \mu X(t)\cdot p((j,k);t) \\ &\quad +(j-1)\alpha \cdot p((j-1,k);t) + (j+1)\beta \cdot p((j+1,k);t) \\ &\quad +\mu X(t)\cdot p((j-1,k);t) + j\mu \cdot p((j,k-1);t). \end{aligned}$$

We use this result to yield the following differential equation:

$$\begin{aligned} \Psi'(y,z;t) &= \frac{\partial \Psi(y,z;t)}{\partial t} \\ &= \sum_{j,k=0}^{\infty} p'((j,k);t) y^j z^k \\ &= \sum_{j,k=0}^{\infty} \Big[-j(\alpha+\beta+\mu)\cdot p((j,k);t) - \mu X(t)\cdot p((j,k);t) \\ &\quad +(j-1)\alpha \cdot p((j-1,k);t) + (j+1)\beta \cdot p((j+1,k);t) \\ &\quad +\mu X(t)\cdot p((j-1,k);t) + j\mu \cdot p((j,k-1);t) \Big] y^j z^k \\ &= (y-1)\mu X(t)\Psi(y,z;t) \\ &\quad +\Big[\mu yz + \alpha y^2 + \beta - (\alpha+\beta+\mu)y\Big] \frac{\partial \Psi(y,z;t)}{\partial y} \end{aligned} \tag{3.11}$$

with initial condition $\Psi(y,z;0) = 1$; an outline of the computations necesary to justify the final equality are provided in the appendix (following the bibliography).

This leads directly to the following expression for $\Psi'(1,0;t)$:

$$\begin{aligned}\Psi'(1,0;t) &= \Big[\alpha+\beta-(\alpha+\beta+\mu)\Big]\frac{\partial\Psi(1,0;t)}{\partial y} \\ &= -\mu\cdot\frac{\partial\Psi(1,0;t)}{\partial y}.\end{aligned}$$

We can write the conditional expectation $\mathrm{E}\left[Y(t) \mid Z(t)=0\right]$ as:

$$\mathrm{E}\left[Y(t) \mid Z(t)=0\right] = \frac{\partial\Psi(1,0;t)}{\partial y} \Big/ \Psi(1,0;t).$$

Next, we easily derive:

$$\begin{aligned}\frac{\partial\Psi(1,0;t)}{\partial y} &= \frac{\partial\Psi(y,z;t)}{\partial y}\Big|_{y=1,z=0} \\ &= \sum_{j,k=0}^{\infty} jy^{j-1}z^k \Pr(y(t)=j, Z(t)=k)\Big|_{y=1,z=0} \\ &= \sum_{j=0}^{\infty} j\Pr(Y(t)=j, Z(t)=0).\end{aligned}$$

Now dividing both sides by $\Psi(1,0;t)$ we get:

$$\begin{aligned}\frac{\partial\Psi(1,0;t)}{\partial y} \Big/ \Psi(1,0;t) &= \sum_{j=0}^{\infty} j\Pr(Y(t)=j, Z(t)=0)/\Psi(1,0;t) \\ &= \sum_{j=0}^{\infty} j\Pr(Y(t)=j, Z(t)=0)/\Pr(Z(t)=0) \\ &= \sum_{j=0}^{\infty} j\Pr(Y(t)=j \mid Z(t)=0) \\ &= \mathrm{E}\left[Y(t) \mid Z(t)=0\right]\end{aligned}$$

Therefore, we can express the hazard function as:

$$\begin{aligned}h(t|\boldsymbol{\theta}) &= -\Psi'(1,0;t)/\Psi(1,0;t) \\ &= \mu\cdot\frac{\partial\Psi(1,0;t)}{\partial y} \Big/ \Psi(1,0;t) \\ &= \mu\cdot\mathrm{E}\left[Y(t)|Z(t)=0\right].\end{aligned}$$

Moolgavkar *et al.* [20] showed that for this model:

$$\begin{aligned}\mathrm{E}\left[Y(t)|Z(t)=0\right] &= \int_0^t \mu X(s)\exp\Big\{\int_0^{t-s}[2\alpha\phi(1,0;u) \\ &\quad -(\alpha+\beta+\mu)]du\Big\}ds,\end{aligned}$$

where $\phi(1,0;u)$ is defined as previously in Section 3.2.2. The second integral in this expression is readily integrated:

$$
\begin{aligned}
& \int_0^{t-s} [2\alpha\phi(1,0;u) - (\alpha+\beta+\mu)]du \\
= & \int_0^{t-s} 2\alpha\phi(1,0;u)du - (\alpha+\beta+\mu)(t-s) \\
= & \; 2\int_0^{t-s} \alpha \cdot \frac{C_1 - C_2[\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}]}{1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}}du - (\alpha+\beta+\mu)(t-s) \\
= & \; 2\int_0^{t-s} \Big\{\alpha C_1\Big[1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}\Big] \\
& +\alpha C_1\Big[\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}\Big] - \alpha C_2\Big[\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}\Big]\Big\} \\
& \times\Big\{\frac{1}{1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}}\Big\}du - (\alpha+\beta+\mu)(t-s) \\
= & \; 2\int_0^{t-s} \Big\{\alpha C_1 - \frac{[-\alpha(C_1-C_2)][\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}]}{1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}}\Big\}du \\
& -(\alpha+\beta+\mu)(t-s) \\
= & \; 2\alpha C_1 u\Big|_0^{t-s} - 2\log\Big\{1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)u\}\Big\}\Big|_0^{t-s} - (\alpha+\beta+\mu)(t-s) \\
= & \; 2\alpha C_1(t-s) - 2\log\Big\{1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)(t-s)\}\Big\} \\
& +2\log\Big\{1-\frac{1-C_1}{1-C_2}\Big\} - (\alpha+\beta+\mu)(t-s) \\
= & \; 2\alpha C_1(t-s) + 2\log\Big\{\frac{1-\frac{1-C_1}{1-C_2}}{1-\frac{1-C_1}{1-C_2}\exp\{\alpha(C_1-C_2)(t-s)\}}\Big\} - (\alpha+\beta+\mu)(t-s) \\
= & \; g(\boldsymbol{\theta};t-s).
\end{aligned}
$$

Thus, the conditional expectation is given by the expression:

$$\mathrm{E}\,[Y(t)|Z(t)=0] = \int_0^t \mu X(s)\exp\{g(\boldsymbol{\theta};t-s)\}ds.$$

Substituting this result into the previous expression for the hazard function we obtain our final expression for the hazard function:

$$h(t|\boldsymbol{\theta}) = \mu^2\int_0^t X(s)\exp\{g(\boldsymbol{\theta};t-s)\}ds, \quad t\geq 0,$$

where $g(\boldsymbol{\theta}; t-s)$ is given above and C_1, C_2 are defined as they were previously in equation (3.9). A deterministic function for the growth of the tissue can now be selected and the remaining integral in the expression for the hazard function can be calculated numerically. I have been using Romberg integration to integrate this function.

3.3.1 Choice of Tissue Growth Function

Both scaled logistic and Gompertz distribution functions have been used in the literature to model the growth of other tissues [18, 20] and thus these have been used as my starting point. Each of these families of distribution functions has several constants that must be chosen in order to get a curve that seems to reasonably fit the pattern we expect for the schwann cell growth. Using a logistic distribution function with three constants to model the tissue growth would give the following form for $X(s)$:

$$X(s) = \frac{K \exp\{\frac{s-a}{b}\}}{1 + \exp\{\frac{s-a}{b}\}}. \tag{3.12}$$

A four-constant Gompertz disribution function would yield:

$$X(s) = K\Big(1 - \exp\{a(1 - \exp\{bs\}) + cs\}\Big).$$

Although both of these families seem to be capable of capturing the shape of tissue growth that is expected from the biological information available to us, I have chosen to use only the logistic family given above. The three constants for this family allow us adequate flexibility to capture any shape of tissue growth that we desire for our modelling purposes.

Choosing the three constants for the logistic family is quite straightforward. The K constant represents the number of cells that the tissue contains in an average adult; in all of our models we assume that the number of tissue cells approaches this constant after a certain age. The remaining constants in the growth function govern the shape of the curve and influence the rate of growth and the age at which the

tissue reaches its maximum size. We will use the notation logistic(K, a, b) to refer a logistic growth function of the form given in (3.12); thus logistic$(10^7, 5.0, 0.8)$ would refer to a logistic growth function with $K = 10^7$, $a = 5.0$, and $b = 0.8$.

Figure 3.1: Three growth functions to be used in model fitting

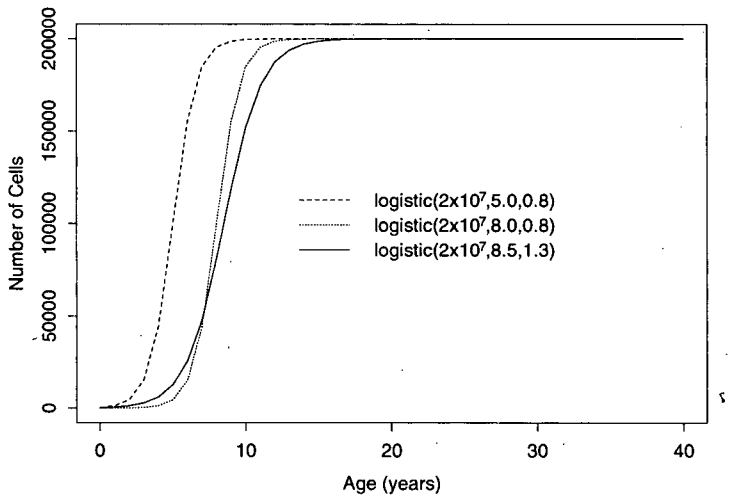


Figure 3.1 is a plot of three different logistic growth functions similar to those that will be used for our data analysis. All three of these functions assume that the maximum number of schwann cells in an adult tissue is 200000. The differences between these functions are the age at which the tissue reaches its maximum size as well as the rate of which the growth occurs. From the figure it is clear that the logistic$(2 \times 10^5, 5.0, 0.8)$ growth function corresponds to a tissue that achieves its maximum size when the individual is approximately age ten; the rate of growth is quite rapid which can be inferred by the steepness of the growth curve. The logistic$(2 \times 10^5, 8.0, 0.8)$ growth function behaves similarly to the logistic$(2 \times 10^5, 5.0, 0.8)$ function in that one curve is essentially a horizontal shift of the other. The biological implications of the logistic$(2 \times 10^5, 8.0, 0.8)$ growth

function are that the tissue experiences a period of very slow growth for the first few years of development. This period is then followed by a spurt of rapid growth with the tissue reaching its maximum size at approximately age 14. The third growth function depicted in figure 3.1 assumes a more gradual rate of tissue growth than the previous growth functions. The logistic$(2 \times 10^5, 8.5, 1.3)$ function assumes that tissue growth is rather slow for the first few years of development, although at a slightly faster rate than the logistic$(2 \times 10^5, 8.0, 0.8)$ function, and approaches its maximum size at about age 17.

Unfortunately, our knowledge of the precise growth of the schwann cells around the vestibular nerve is quite limited. From the limited information available to us, any of these functions could very well be a feasible growth function to model the growth of the tissue. The information that is perhaps most important for our modelling, namely the number of cells present in an adult tissue, is particularly difficult to obtain. We have estimated the upper bound on this quantity to be approximately 10^7; we are however unable to provide an assessment of the reliability of this estimate. We have therefore decided to fit our models using a variety of different growth functions and assess the sensitivity of the model fit and parameter estimates to the growth function selected.

3.3.2 Likelihood Construction

Given an expression for the hazard function $h(t|\boldsymbol{\theta})$ we can express the density function for the time to the onset of the first VS as:

$$f(t|\boldsymbol{\theta}) = h(t|\boldsymbol{\theta}) \exp\Big\{ - \int_0^t h(u|\boldsymbol{\theta}) du \Big\}, \quad t \geq 0.$$

If our data consist of n individuals with onset times $t_1, ..., t_n$, and assuming that individuals are independent, the likelihood for the data can be written as a product

of density functions:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\left(h(t_i|\boldsymbol{\theta})\exp\Big\{-\int_0^{t_i}h(u|\boldsymbol{\theta})du\Big\}\right) \\ &= \left(\prod_{i=1}^{n}h(t_i|\boldsymbol{\theta})\right)\left(\prod_{i=1}^{n}\exp\Big\{-\int_0^{t_i}h(u|\boldsymbol{\theta})du\Big\}\right) \\ &= \left(\prod_{i=1}^{n}h(t_i|\boldsymbol{\theta})\right)\exp\Big\{-\sum_{i=1}^{n}\int_0^{t_i}h(u|\boldsymbol{\theta})du\Big\} \end{aligned}$$

The log-likelihood, $l(\boldsymbol{\theta})$, can then be written as:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\{h(t_i|\boldsymbol{\theta})\} - \sum_{i=1}^{n}\int_0^{t_i}h(u|\boldsymbol{\theta})du$$

The log-likelihood is then maximized with respect to $\boldsymbol{\theta}$ to obtain the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Given this estimate of the parameter vector, it is possible to compute an estimate of the probability that a patient develops a tumour before a specific time t. An estimate of this probability, $F(t|\hat{\boldsymbol{\theta}})$, is computed as follows:

$$F(t|\hat{\boldsymbol{\theta}}) = \widehat{\Pr}(T \le t) = 1 - \exp\Big\{-\int_0^{t}h(u|\hat{\boldsymbol{\theta}})du\Big\}, \quad t \ge 0.$$

3.3.3 Genotype Information

It may also be of interest to include information about the type of mutation that a patient has sustained to their *NF2* gene into the model. We may wish to include this information in order to obtain probability predictions for individuals bearing a certain type of genetic mutation. The most obvious way of incorporating this information into the model is to allow the vector of parameters for the model to be different for patients with different types of mutations. Suppose a patient can have one of k different types of mutations and the vector of model parameters for an individual with mutation type j is denoted by $\boldsymbol{\theta}_j$. Let m_i denote the mutation type for patient i; thus, the data that are collected on patient i is the vector (t_i, m_i) and patient i's contribution to the likelihood will be:

$$f(t_i|\boldsymbol{\theta}_{m_i}) = h(t_i|\boldsymbol{\theta}_{m_i})\exp\Big\{-\int_0^{t_i}h(u|\boldsymbol{\theta}_{m_i})du\Big\}, \quad t_i \ge 0.$$

Additionally, let D_j be the set of patients in our data set with mutation type j. Then the likelihood for the data can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k) &= \prod_{i=1}^{n} \left(h(t_i|\boldsymbol{\theta}_{m_i}) \exp\Big\{ - \int_0^{t_i} h(u|\boldsymbol{\theta}_{m_i}) du \Big\} \right) \\ &= \prod_{j=1}^{k} \left(\prod_{i \in Dj} h(t_i|\boldsymbol{\theta}_j) \exp\Big\{ - \int_0^{t_i} h(u|\boldsymbol{\theta}_j) du \Big\} \right) \\ &= \left(\prod_{j=1}^{k} \prod_{i \in D_j} h(t_i|\boldsymbol{\theta}_j) \right) \exp\Big\{ - \sum_{j=1}^{k} \sum_{i \in D_j} \int_0^{t_i} h(u|\boldsymbol{\theta}_j) \Big\} du. \end{aligned}$$

The log-likelihood can therefore be written as:

$$l(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k) = \sum_{j=1}^{k} \sum_{i \in D_j} \log\big\{ h(t_i|\boldsymbol{\theta}_j) \big\} - \sum_{j=1}^{k} \sum_{i \in D_j} \int_0^{t_i} h(u|\boldsymbol{\theta}_j) du$$

Again, estimates of the parameter vectors $\boldsymbol{\theta}_j$ $(j = 1, ..., k)$ can be obtained by maximizing the log-likelihood with respect to the parameters. An obvious concern with this approach is that allowing the model parameters to differ for each mutation type greatly increases the number of parameters that must be estimated from the data; for large data sets this might not be a serious concern. However, since NF2 data sets are typically small and genotype information is often missing from the patient records, it is desirable to reduce the number of parameters to as few as are feasible. Constraints can be imposed to reduce the number of parameters and this is discussed in some detail below.

Upon obtaining estimates of the model parameters $\hat{\boldsymbol{\theta}}_j, j = 1, ..., k$, it is possible to estimate quantities of interest. The probability that an individual with mutation type j will develop a tumour by age t, $F(t|\boldsymbol{\theta}_j)$, will be denoted by $F_j(t)$ and can be estimated as:

$$\widehat{F_j(t)} = 1 - \exp\Big\{ - \int_0^{t} h(u|\hat{\boldsymbol{\theta}}_j) du \Big\}, \quad t \geq 0.$$

Additionally, we may also wish to compare different genotypes using these mutation models. Two genotypes could be compared by estimating the risk of developing a

tumour at a given age t for an individual with one genotype relative to the risk for an individual with another genotype. The risk of developing a tumour at age t for an individual with mutation type i relative to an individual with mutation type j is denoted by $r_{ij}(t)$ and can be estimated as:

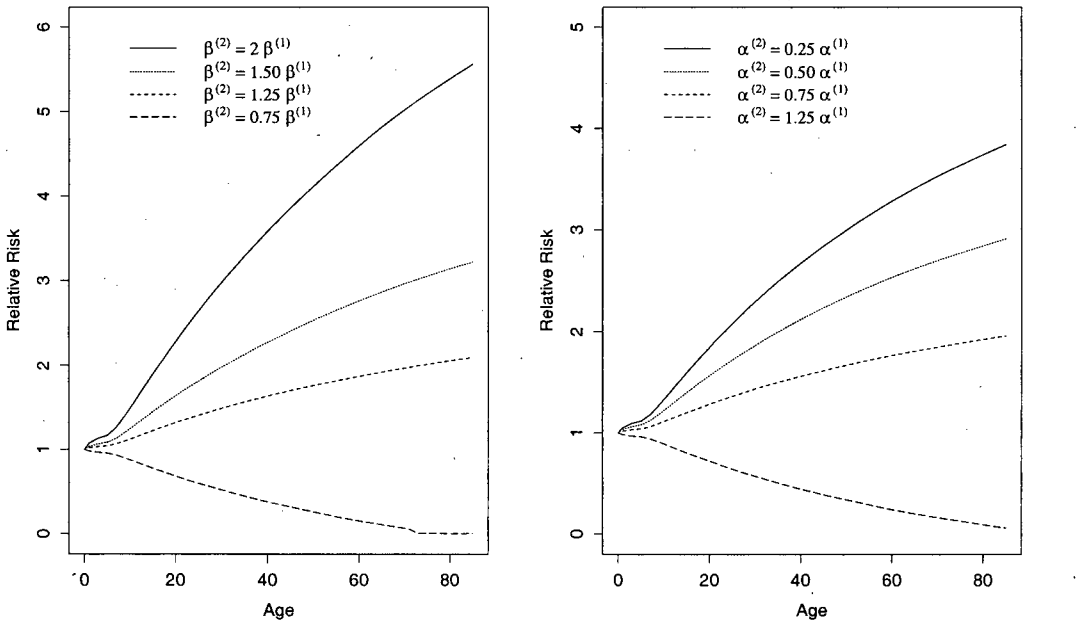
$$\widehat{r_{ij}(t)} = \frac{h(t|\hat{\boldsymbol{\theta}}_i)}{h(t|\hat{\boldsymbol{\theta}}_j)}. \tag{3.13}$$

Clearly the estimates for the relative risks given in equation (3.13) have a dependence on t; this is interpretable as a patient's risk of developing a VS will change with their age. In general, the hazard functions used in equation (3.13) to estimate the relative risk need not be evaluated at the same age. As an example, suppose there is interest in comparing the risk of developing a VS for an individual with genotype i at age t, relative to an individual with genotype j at age 10, then this relative risk could be estimated by $h(t|\hat{\boldsymbol{\theta}}_i)/h(10|\hat{\boldsymbol{\theta}}_j)$. For our analyses we will estimate the relative risks according to equation (3.13) by estimating the risk of developing a VS for an individual with one genotype at age t, relative to the risk of developing a VS for an individual with another genotype at the same age.

Suppose the model under consideration has three rate parameters for each mutation type; these would be the growth, death, and mutation rates for the intermediate cells. We can denote the parameter vector for the jth mutation type as $\boldsymbol{\theta}_j = (\alpha^{(j)}, \beta^{(j)}, \mu^{(j)})$ where $\alpha^{(j)}, \beta^{(j)}$, and $\mu^{(j)}$ are the growth, death, and mutation rates respectively. The superscripts are used in place of subscripts to identify the mutation type as subscripts on these parameters are typically used in the literature to denote the stage of the model under discussion. A constraint that can be employed to reduce the number of model parameters is $\mu^{(j)} = \mu$ $(j = 1, ..., k)$; this would imply that the mutation rates for the chance mutations are equal for patients with different mutations of the *NF2* gene. This model assumes that the type of mutation of the *NF2* gene affects only the rates at which intermediate cells in the tissue divide and die. The number of parameters in the full and reduced models are $3k$ and $2k + 1$ respectively, where k is the number of different mutation types in

the data set. Clearly, the necessity to reduce the number of model parameters will depend on both the number of mutation types to be considered in the analysis and more importantly on the interpretability of the constraint given above.

Figure 3.2: Plots of relative risks versus age obtained from a 3-hit model with specifically chosen parameter values: $\boldsymbol{\theta}_1 = (\alpha^{(1)}, \beta^{(1)}, \mu^{(1)}) = (1.73 \times 10^{-1}, 1.93 \times 10^{-1}, 4.10 \times 10^{-4})$; values for the components of $\boldsymbol{\theta}_2$ vary across the plots.



(a) Assuming equal mutation and growth rates; plots are labelled by the death rate used

(b) Assuming equal mutation and death rates; plots are labelled by the growth rate used

Figure 3.2 is an illustration of the estimated relative risks described above. The plots in this figure were produced by fixing the parameters of the 3-hit model to suitable values and plotting the relative risk as a function of age. For these plots the mutation rates for the two genotypes are assumed equal and the growth and death rates for the intermediate cells are allowed to differ across mutation types. In the left panel of the figure the growth rates for the two mutation types are assumed

to be equal and the relative risk is estimated, as a function of age, for a range of death rates. The plot suggests that individuals with mutation type 1 have a higher risk of developing a VS than do individuals with mutation type 2 when the death rate for their intermediate cells is lower than for individuals with the latter type of mutation; additionally, this relative risk increases as a function of age. This is quite reasonable as one would expect that when the intermediate cells die more quickly there would be fewer of them present in the tissue to sustain the final mutation and thus the likelihood of a tumour cell developing in the tissue would be smaller. The right panel of the figure shows a plot very similar to the plot from the left panel except that the death rates for the intermediate cells are assumed to be equal across the two genotypes and here the growth rates are allowed to vary across genotypes. Individuals with mutation type 1 clearly have a higher risk of developing a VS than individuals with mutation type 2 when the growth rate for their intermediate cells is higher than individuals bearing a mutation of the latter type; as previously this relative risk is an increasing function of age for the chosen parameter values. Again, this is quite reasonable as a more rapid growth of intermediate cells would imply that there are more cells present in the tissue that need only to acquire the final chance mutation to develop into tumour cells. It is possible to produce plots similar to these for various parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$; the plots included here were meant only as examples to illustrate the merits of using relative risks to summarize the models fit using genotype information.

Chapter 4

Estimating the Age at Onset of the Second VS

Estimating the age at which a patient with NF2 will develop both the left and right VSs is also an interesting problem. Estimation of this quantity would require data related to both the onset of the left and right VSs; previously we have been concerned with only the age at onset of the first VS. A subset of the MUK data has information recorded on the onset of both VSs for 144 patients and are suitable for our modelling purposes. An important note concerning these data is that the information recorded for each patient is the age at which these tumours were detected, not necessarily the age at onset for the two tumours; this point will be relevant in the interpretation of our estimates.

The ages at which the right and left VSs develop are quite likely correlated on an individual and thus we choose to model these times as bivariate random variables. To allow for such dependence, we employ the the Kimeldorf and Sampson family of copulas [11, 10] and specify the margins using the appropriate multi-hit model. This model is sometimes called the Clayton-Oakes model [3, 24] and is often used to model bivariate survival data.

Let $T_i^{(l)}$ and $T_i^{(r)}$ represent the ages at onset of the left and right VSs respec-

tively for individual i. We assume that these two random variables have marginal distribution functions F_l and F_r respectively; the marginal densities are denoted by f_l and f_r. The notation $C(u, v; \delta)$ is used to represent the Kimeldorf and Sampson family of copulas, with dependence parameter δ. The specific form of this family is given as:

$$C(u, v; \delta) = (u^{-\delta} + v^{-\delta} - 1)^{-1/\delta}, \quad 0 \leq \delta < \infty,$$

where U and V are random variables with Uniform(0,1) marginal distributions. The joint density for U and V under such a model is given by:

$$c(u, v; \delta) = (1 + \delta)[uv]^{-\delta-1}(u^{-\delta} + v^{-\delta} - 1)^{-2-1/\delta}, \quad 0 \leq \delta < \infty.$$

Thus, we replace u and v in these expressions by $F_l(t^{(l)})$ and $F_r(t^{(r)})$ for our application. The joint density function of $T^{(l)}$ and $T^{(r)}$ can now be written as:

$$\begin{aligned} f_{l,r}(t^{(l)}, t^{(r)}) &= \frac{\partial^2 C(F_l(t^{(l)}), F_r(t^{(r)}); \delta)}{\partial F_l \partial F_r} \Big(\frac{\partial F_l(t^{(l)})}{\partial t^{(l)}}\Big)\Big(\frac{\partial F_r(t^{(r)})}{\partial t^{(r)}}\Big) \\ &= c(F_l(t^{(l)}), F_r(t^{(r)}); \delta) f_l(t^{(l)}) f_r(t^{(r)}), \quad 0 \leq \delta < \infty. \end{aligned}$$

The marginal distributions for $T^{(l)}$ and $T^{(r)}$ are assumed to be identical; we will denote the hazard, density, and distribution functions for these random variables as $h(t|\boldsymbol{\theta}), f(t|\boldsymbol{\theta})$, and $F(t|\boldsymbol{\theta})$ respectively. This assumption is justifiable as there is no reason to assume that a VS on one side of the head is more likely to develop by a certain age than a VS on the other side of the head. This assumption simplifies the expression for the joint density of $T^{(l)}$ and $T^{(r)}$:

$$\begin{aligned} f_{l,r}(t^{(l)}, t^{(r)}) &= c(F(t^{(l)}|\boldsymbol{\theta}), F(t^{(r)}|\boldsymbol{\theta}); \delta) f(t^{(l)}|\boldsymbol{\theta}) f(t^{(r)}|\boldsymbol{\theta}) \\ &= (1 + \delta)\Big(F(t^{(l)}|\boldsymbol{\theta})^{-\delta} + F(t^{(r)}|\boldsymbol{\theta})^{-\delta} - 1\Big)^{-2-1/\delta} \\ &\quad \times \Big[F(t^{(l)}|\boldsymbol{\theta}) F(t^{(r)}|\boldsymbol{\theta})\Big]^{-\delta-1} f(t^{(l)}|\boldsymbol{\theta}) f(t^{(r)}|\boldsymbol{\theta}), \quad 0 \leq \delta < \infty. \end{aligned} \tag{4.1}$$

In the previous sections we derived an expression for the hazard function for the time until the onset of the first VS. The hazard function derived was the hazard for the entire tissue at risk of developing the VS; this tissue consisted of all of the

schwann cells surrounding both the left and right vestibular nerves. In this section, the hazard function given in our expressions will denote the hazard function for the time until the generation of the first tumour cell in a single vestibular nerve; this hazard function will specify the marginal distribution of both $T^{(l)}$ and $T^{(r)}$. The derivation of the hazard function proceeds identically to the derivation from the previous sections, except that the number of cells that are assumed to be initially present in the tissue are divided in half. Using previous results, and the relationships between the hazard, distribution, and density functions, we can reexpress the joint density from equation (4.1) in terms of this hazard function:

$$\begin{aligned}
f_{l,r}(t^{(l)}, t^{(r)}) &= (1+\delta) \prod_{j\in\{l,r\}} \Big[1 - \exp\Big\{-\int_0^{t^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}\Big]^{-\delta-1} \\
&\quad \times \Big[-1 + \sum_{j\in\{l,r\}} \Big(1 - \exp\Big\{-\int_0^{t^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}\Big)^{-\delta}\Big]^{-2-1/\delta} \\
&\quad \times h(t^{(l)}|\boldsymbol{\theta}) h(t^{(r)}|\boldsymbol{\theta}) \exp\Big\{-\int_0^{t^{(l)}} h(s|\boldsymbol{\theta})ds\Big\} \exp\Big\{-\int_0^{t^{(r)}} h(s|\boldsymbol{\theta})ds\Big\} \\
&= (1+\delta) \prod_{j\in\{l,r\}} \Big[1 - \exp\Big\{-\int_0^{t^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}\Big]^{-\delta-1} \\
&\quad \times \Big[-1 + \sum_{j\in\{l,r\}} \Big(1 - \exp\Big\{-\int_0^{t^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}\Big)^{-\delta}\Big]^{-2-1/\delta} \\
&\quad \times h(t^{(l)}|\boldsymbol{\theta}) h(t^{(r)}|\boldsymbol{\theta}) \exp\Big\{-\sum_{j\in\{l,r\}} \int_0^{t^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}
\end{aligned}$$

If our data consist of n patients, with observed onset times $(t_1^{(l)}, t_1^{(r)}), ..., (t_n^{(l)}, t_n^{(r)})$, then our log-likelihood is simply:

$$\begin{aligned}
l(\boldsymbol{\theta}) &= n\log(1+\delta) - (1+\delta) \sum_{i=1}^{n} \sum_{j\in\{l,r\}} \log\Big\{1 - \exp\Big\{\int_0^{t_i^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}\Big\} \\
&\quad -(2+1/\delta) \sum_{i=1}^{n} \log\Big\{-1 + \sum_{j\in\{l,r\}} \Big(1 - \exp\Big\{\int_0^{t_i^{(j)}} h(s|\boldsymbol{\theta})ds\Big\}\Big)^{-\delta}\Big\} \\
&\quad + \sum_{i=1}^{n} \sum_{j\in\{l,r\}} \log\Big\{h(t_i^{(j)}|\boldsymbol{\theta})\Big\} - \sum_{i=1}^{n} \sum_{j\in\{l,r\}} \int_0^{t_i^{(j)}} h(s|\boldsymbol{\theta})ds. \qquad (4.2)
\end{aligned}$$

The log-likelihood given in equation (4.2) can be maximized with respect to the parameters $\boldsymbol{\theta}$ and δ to obtain parameter estimates. These estimates can be imputed into our expression for the bivariate distribution function of $T^{(l)}$ and $T^{(r)}$, denoted by $F_{l,r}(t^{(l)}, t^{(r)})$, to compute estimates of the probability that an individual will develop both tumours by a given age; note that $F_{l,r}(t^{(l)}, t^{(r)}) = C(F(t^{(l)}|\boldsymbol{\theta}), F(t^{(r)}|\boldsymbol{\theta}); \delta)$. A plot of $\widehat{F_{l,r}(t,t)}$ versus t can be constructed and compared to a plot of the empirical bivariate distribution function to assess the fit of the model. The bivariate empirical distribution function would be computed as follows:

$$\widehat{F_E(t,t)} = \frac{\#\{T^{(l)} \leq t, T^{(r)} \leq t\}}{n} = \frac{\#\{\max(T^{(l)}, T^{(r)}) \leq t\}}{n}.$$

An interesting feature of this model is that it allows us to estimate the association between the ages at onset of the left and right VSs. This association is characterized by the model parameter δ. Interpreting the magnitude of an estimate of this parameter will be discussed in Chapter 5 and further details can be found in [10].

Chapter 5

Results

5.1 Knudson's Model

Knudson's model was fit using the age at onset of the first VS variable from both the MUK and SPOR datasets. The fitting of the model is fairly simple to implement and requires little computational effort. There is a single time-dependent parameter that must be estimated from the data at several pre-selected time points. The estimation is performed using a non-linear weighted least-squares procedure described previously in Chapter 2. All computations were performed using programs written in the C programming language.

For our fitting eight time points were selected at which the model parameter would be estimated. Recall that the model parameter represents the fraction of cell divisions that have occurred prior to time t and is denoted by $d(t)$. There are two other quantities that must be specified prior to the fitting of the model. These are the expected number of tumours eventually acquired by an NF2 patient, denoted by $m(\infty)$, and the ratio of the expected number of cell divisions in an individual's life to the number of schwann cells originally present in the tissue; this latter quantity will be denoted by $a(\infty)/b(0)$. The results presented here have been generated using $m(\infty) = 2$ and $a(\infty)/b(0) = 2 \times 10^6$. It is natural to specify the expected number of

tumours eventually acquired by an NF2 patient to be two as all NF2 patients used in our fitting have bilateral VS. The second quantity is more difficult to justify as we require some information about the number of cells present in the tissue and the number of cellular divisions that occur throughout an individual's life. The value chosen is identical to the value used by Hethcote and Knudson [8] in their application to retinoblastoma and is also consistent with the growth functions for the tissue that we assume in the next section. As well, the results obtained in the model fitting depend very little on the value chosen for this quantity. Hethcote and Knudson [8] report that for their application, estimates of $d(t)$ change by less than 1% if any value between 10^5 and 10^8 is selected for $a(\infty)/b(0)$. Our experiences with perturbations of this quantity are consistent with those described by the aforementioned authors.

Table 5.1 shows the time points that were selected for the estimation of the fraction of cell divisions and the resulting estimates. Note that estimates of the standard errors are not provided as the method of estimation does not suggest an obvious method for computing standard errors. These estimates can be imputed into equations (2.1) and (2.2) to yield estimates of the cumulative probability that an individual will develop a VS prior to a certain age; these estimates are computed for both NF2 and sporadic patients. A plot of these model estimated probabilities is given in Figure 5.1; plots of the empirical distribution functions have also been added to assess the fit of the model. Overall, the model fit appears to be adequate despite a few deviations between the observed and fitted incidence functions. The model tends to under-predict the incidence for NF2 patients for a range of ages (ages 35-60); it also over-predicts the incidence for the sporadic cases on a range of ages (ages 25-50). A chi-square goodness of fit test was used to test the adequacy of the model fit and suggested that there is some evidence of departure between the model and the data ($\chi^2_6 = 10.346$, p-value $= 0.066$). To compute the test statistic, observed and expected incidences of VS were compared in several age intervals; intervals with observed or expected counts of less than five were combined

Figure 5.1: Plots of empirical and estimated probabilities from Knudson's model for both NF2 patients and sporadic cases

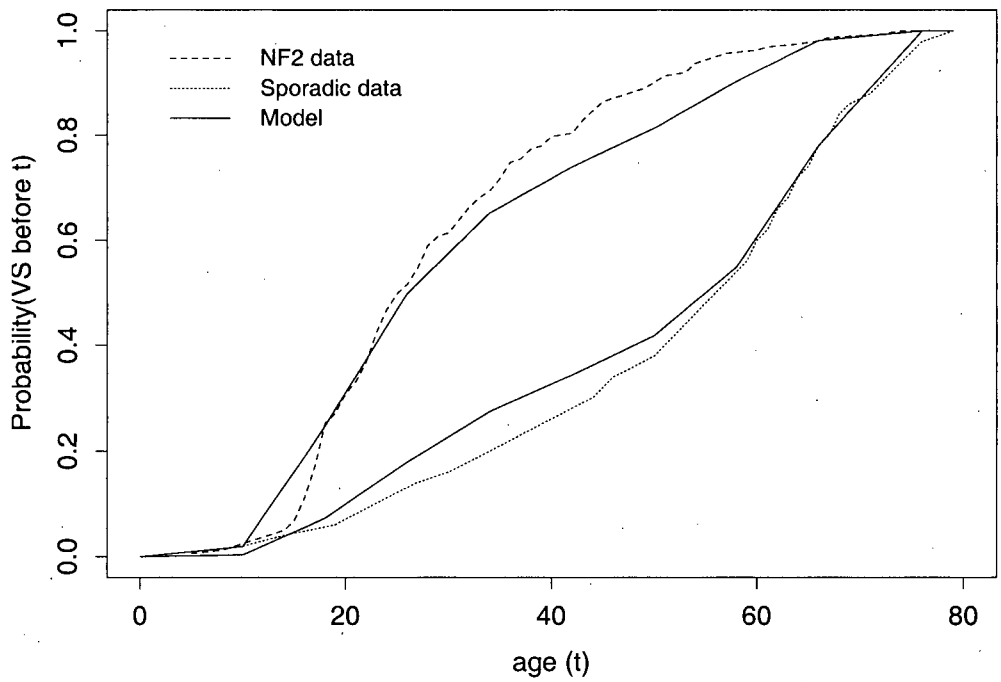


Table 5.1: Estimates of the fraction of cell divisions $d(t)$ at various time points from Knudson's model

age (t)	0	10	18	26	34	42	50	58	66	∞
$\widehat{d(t)}$	0	0.006	0.087	0.204	0.300	0.371	0.445	0.572	0.794	1

for asymptotic considerations. These results have motivated us to further explore the two-mutation hypothesis with other models to see if results are consistent across different models.

5.2 2-hit Models

5.2.1 Deterministic Tissue Growth

The model fitting for the 2-hit model with deterministic growth of the tissue is quite simple. There is no need to employ a numerical routine to perform the maximization of the likelihood function given in equation (3.1), as a closed form expression for the mutation rate parameter exists. We have only to select a suitable growth function for the tissue and input this function into equation (3.2) to compute our estimate of the mutation rate.

Three candidate growth functions were selected for the model fitting. These are the same three functions from the logistic family that are plotted in Figure 3.1. It is assumed in each case that the tissue originally contains 20 cells and after a certain age reaches a constant size of 10^7 cells; as was discussed previously, the three growth functions differ from one another in the age at which the tissue reaches its maximum size and the rate at which the tissue grows. In computing the mutation rate estimate, there is a need to integrate the growth function several times, each time over a finite interval; these integrals have been computed numerically using Romberg integration [16]. All computations were implemented using programs written in the C programming language.

The models were fit to both the age at onset of the first VS and age at onset of hearing loss variables from the Manchester data set, as well as the age at onset of hearing loss variable from the multi-source data set. Parameter estimates and estimated standard errors from the model fittings are summarized in Tables 5.2, 5.3, and 5.4 for the age at onset of first VS (MUK), age at onset of hearing loss (MUK) and age at onset of hearing loss (FSS) respectively. The estimates for the mutation rates obtained using the three different growth functions are ordered identically across the three different data sets; specifically, the logistic$(10^7, 5.0, 0.8)$ growth function consistently yields the smallest rate estimate, while the logistic$(10^7, 8.5, 1.3)$ function produces the highest estimate of the mutation rate for all three data sets. It is important to recognize that the magnitude of the mutation rate estimate for this model depends heavily on the number of tissue cells that are assumed to be present in the adult tissue. This is quite clear from the expression for the mutation rate estimate. If the number of tissue cells in an adult tissue was rescaled by a multiplicative factor of k then the estimate of the mutation rate would consequently be rescaled by a factor of $1/k$. Model fitted cumulative distribution functions for the age at onset random variable will not however be affected by rescaling the tissue growth function by a multiplicative factor. For example, models fit using the logistic$(10^7, 8.5, 1.3)$ and logistic$(10^5, 8.5, 1.3)$ growth functions will produce identical estimates of the cumulative distribution function despite having different estimates for the mutation rate.

Estimates of the cumulative distribution functions obtained using the three different data sets are presented in Figures 5.2, 5.3 and 5.4. In each of these figures, the three estimated distribution functions, each obtained using a different growth function for the tissue, are plotted against the empirical distribution function. For all three data sets it is evident that models fit assuming logistic$(10^7, 8.0, 0.8)$ and logistic$(10^7, 8.5, 1.3)$ growth functions yield very similar estimates of the cumulative distribution function. Additionally, these estimates are also more consistent

Table 5.2: Model fitting results for 2-hit model with deterministic tissue growth using MUK data

Variable: Age at onset of the first VS		

Growth function: logistic(10^7,5.0,0.8)		
Parameter	Estimate	Estimated SE
μ	4.052×10^{-9}	0.317×10^{-9}
value of log-likelihood: -686.280		

Growth function: logistic(10^7,8.0,0.8)		
Parameter	Estimate	Estimated SE
μ	4.609×10^{-9}	0.361×10^{-9}
value of log-likelihood: -669.359		

Growth function: logistic(10^7,8.5,1.3)		
Parameter	Estimate	Estimated SE
μ	4.715×10^{-9}	0.369×10^{-9}
value of log-likelihood: -665.474		

with the empirical distribution function than the estimate obtained assuming a logistic$(10^7, 5.0, 0.8)$ model for the growth of the tissue. Overall, the fit of these simple one-parameter models to the empirical distribution functions is fair. In particular, all three of the models tend to predict an earlier onset of the tumours than is reflected in our data. The explanation for the lack of model fit could be any of several things. One explanation could be that our assumed growth functions are all incorrect; this possibility is difficult to assess given our limited information on the growth of the tissue. Uncertainty in our assumptions on the growth of the tissue is one motivation for choosing a model where the parameters governing the growth of the tissue are estimated from the data. The 2-hit model with stochastic growth of the tissue discussed previously is an example of such a model. Results from the fitting of such a model will be discussed in the next section.

Table 5.3: Model fitting results for 2-hit model with deterministic tissue growth using MUK data

Variable: Age at onset of hearing loss		
Growth function: logistic(10^7,5.0,0.8)		
Parameter	Estimate	Estimated SE
μ	4.524×10^{-9}	0.377×10^{-9}
value of log-likelihood: -590.127		

Growth function: logistic(10^7,8.0,0.8)		
Parameter	Estimate	Estimated SE
μ	5.228×10^{-9}	0.436×10^{-9}
value of log-likelihood: -573.221		

Growth function: logistic(10^7,8.5,1.3)		
Parameter	Estimate	Estimated SE
μ	5.364×10^{-9}	0.447×10^{-9}
value of log-likelihood: -569.611		

Table 5.4: Model fitting results for 2-hit model with deterministic tissue growth using FSS data

Variable: Age at onset of hearing loss		
Growth function: logistic(10^7,5.0,0.8)		
Parameter	Estimate	Estimated SE
μ	4.625×10^{-9}	0.358×10^{-9}
value of log-likelihood: -685.683		

Growth function: logistic(10^7,8.0,0.8)		
Parameter	Estimate	Estimated SE
μ	5.358×10^{-9}	0.415×10^{-9}
value of log-likelihood: -669.483		

Growth function: logistic(10^7,8.5,1.3)		
Parameter	Estimate	Estimated SE
μ	5.499×10^{-9}	0.426×10^{-9}
value of log-likelihood: -662.696		

Figure 5.2: Plots of empirical and model predicted probabilities for 2-hit model with deterministic growth of tissue; using age at onset of first VS (MUK data)

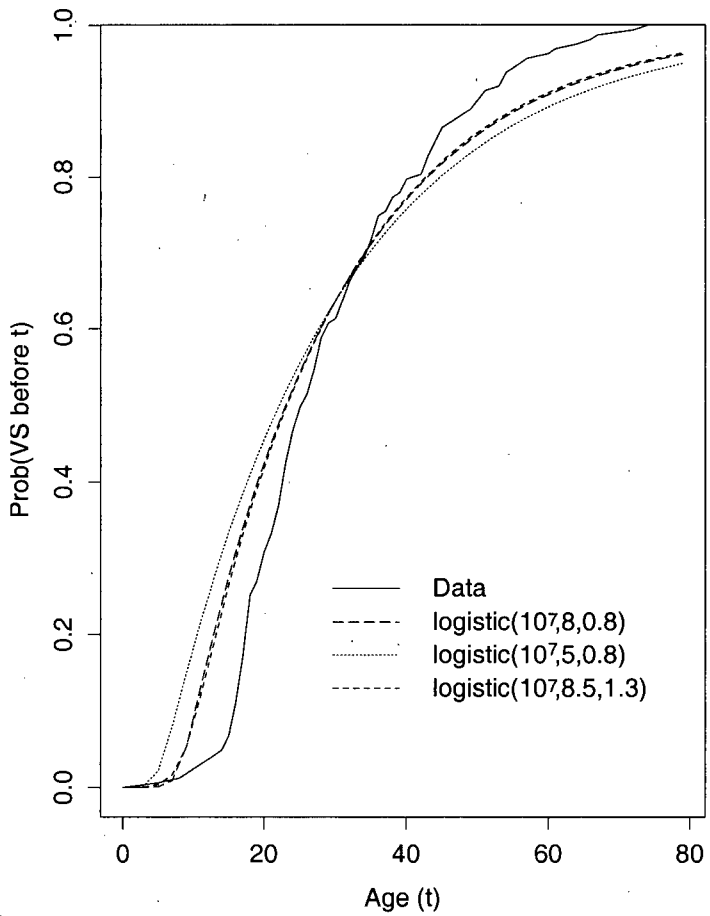


Figure 5.3: Plots of empirical and model predicted probabilities for 2-hit model with deterministic growth of tissue; using age at onset of hearing loss (MUK data)

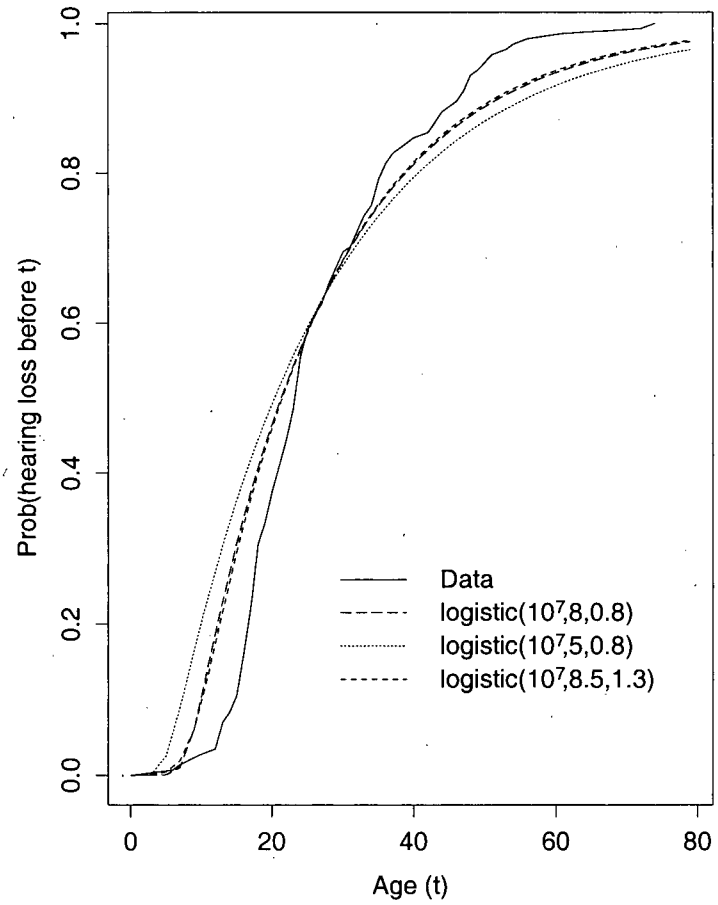
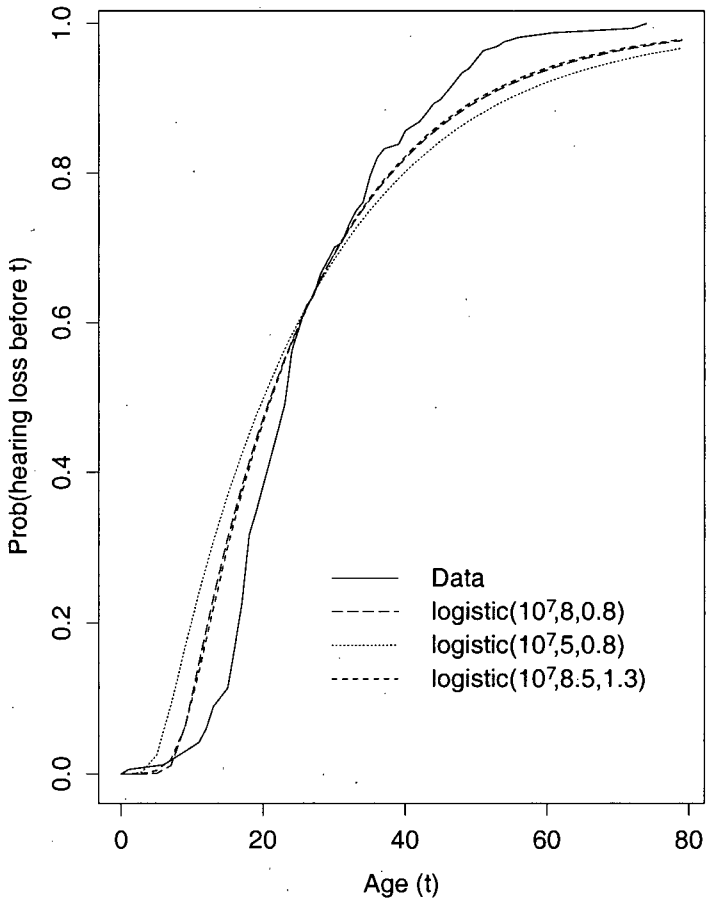


Figure 5.4: Plots of empirical and model predicted probabilities for 2-hit model with deterministic growth of tissue; using age at onset of hearing loss (FSS data)



5.2.2 Stochastic Tissue Growth

Fitting the 2-hit model with stochastic tissue growth is more complicated than fitting the model from the previous section. One of the most obvious differences between these two models with respect to fitting, is that the parameters from the fully stochastic model do not have closed form maximum likelihood estimates; all three parameters must be estimated by a numerical routine. For this model, a likelihood for the data was constructed according to the method described in section 3.3.2 by expressing the likelihood function in terms of the hazard function for the time to the first tumour cell. The log-likelihood was maximized using a Quasi-Newton routine [23] and all computations were implemented using programs written in C. All definite integrals in the expression for the likelihood were computed numerically using Romberg integration [16].

The growth and death of the tissue cells in this model are governed by a stochastic process and hence there is no need to select a growth function for the tissue. We do however need to select the number of cells initially present in the tissue, denoted previously by N. We have chosen to use $N = 20$ for our fitting; this is consistent with the number of tissue cells initially present for the model with deterministic growth of the tissue. Tables 5.5, 5.6 and 5.7 contain the parameter estimates and the estimated standard errors from the fitting of the model to the three data sets. Estimates of the model parameters are consistent in magnitude across the three data sets. The ordering of the growth, death and mutation rates is also consistent across the three data sets. Several different vectors of starting values were selected for the Quasi-Newton routine; most of these yielded consistent solutions. Several starting values resulted in the convergence to a local maximum; the value of the log-likelihood was used to discriminate between local and global maxima. The value of the log-likelihood evaluated at the maximum likelihood estimates of the model parameters is given in the aforementioned tables below the parameter estimates.

Table 5.5: Model fitting results for 2-hit model with stochastic tissue growth using MUK data

Variable: Age at onset of the first VS		
Parameter	Estimate	Estimated SE
α	4.214×10^{-1}	1.010×10^{-1}
β	3.352×10^{-1}	1.013×10^{-1}
μ	3.154×10^{-4}	0.642×10^{-4}
value of log-likelihood: -659.966		

Table 5.6: Model fitting results for 2-hit model with stochastic tissue growth using MUK data

Variable: Age at onset of hearing loss		
Parameter	Estimate	Estimated SE
α	4.999×10^{-1}	1.209×10^{-1}
β	4.009×10^{-1}	1.112×10^{-1}
μ	3.250×10^{-4}	0.714×10^{-4}
value of log-likelihood: -568.175		

Table 5.7: Model fitting results for 2-hit model with stochastic tissue growth using FSS data

Variable: Age at onset of hearing loss		
Parameter	Estimate	Estimated SE
α	4.844×10^{-1}	1.138×10^{-1}
β	3.860×10^{-1}	1.049×10^{-1}
μ	3.406×10^{-4}	0.771×10^{-4}
value of log-likelihood: -656.965		

Upon obtaining estimates of the model parameters, it is possible to plot the estimate of the cumulative distribution function for the age at onset variable. A plot of the empirical distribution function can be compared to these model fitted cumulative distribution functions to assess the overall fit of the model. Figures 5.5, 5.6 and 5.7 are plots of the fitted cumulative distribution function versus the empirical distribution function for the three data sets. The fit of this model is clearly an improvement over the model that assumes a deterministic function for the growth of the tissue. For all three data sets the model tends to over-predict the incidence of tumours at young ages (ages less than 15 years); this does not appear to be a serious concern however, as the model seems to fit reasonably well overall.

As an additional check on the fit of the model, it is possible to simulate onset times from the fitted model and compare their distribution to that of our data. Simulating onset times from the fitted model, which we will denote by T^*, can be done according to the following simple algorithm:

- Recall that under our model $F(T|\boldsymbol{\theta}) = 1 - \exp\Big\{-\int_0^T h(s|\boldsymbol{\theta})ds\Big\} \sim$ Uniform(0,1).
- Generate a Uniform(0,1) random variable U^*.
- Set $U^* = 1 - \exp\Big\{-\int_0^{T^*} h(s|\boldsymbol{\theta})ds\Big\}$ and solve for the onset time T^*.

Although it would appear that simulating onset times is quite a simple task, the final step in the algorithm does require the use of a numerical method to solve the given equality. I have used the bisection method for this and have found it to be quite successful. Using the bisection method for this requires that we bracket the onset time between two points; I chose to use the interval $[0, 100]$ and have encountered no difficulties with finding a sensible solution thus far.

Samples of onset times were simulated from each of the three fitted models. The number of onset times simulated from each model was identical to the sample size of patients used to fit the model. Histograms of the simulated onset times and

of the actual data are presented in Figures 5.8, 5.9 and 5.10 for the three data sets. The histograms for the simulated onset times are quite similar in their distribution to the actual data for onset times greater than 20; the two distributions do however differ for onset times less than 20. Simulations from the fitted models generate a higher number of early onset times (onset times less than 10) than are represented in our data sets. The quartiles for the simulated data sets match the quartiles of the actual data very closely; in fact the medians for the simulated and actual times are identical for all three of the data sets.

5.2.3 Genotype

Patients used for the fitting of this model were stratified according to mutation type into one of two groups: patients with protein truncating mutations and patients with other known mutation type. The group of patients with protein truncating mutations includes the patients from our dataset with either frameshift or nonsense mutations; these types of mutations produce a similar effect on the protein product and thus have been grouped together. Our second stratum includes all other known mutation types from our dataset; this group will be less homogeneous than the protein truncating group with respect to the effects on protein product produced by different mutation types in this stratum. This stratification was chosen as a result of the small number of patients in our dataset with identified mutation type. For simplicity we will refer to the patients with truncating mutations as having genotype 1 and patients with other types of mutations as having genotype 2. Two models were fit to the data: the first assuming that the mutation rate parameters for the two genotypes were identical; and the second model allowing these mutation rate parameters to be different across the two genotypes. Both of these models were fit to the FSS data set on the age at onset of hearing loss variable.

Parameter estimates and estimated standard errors from the five-parameter model are given in Table 5.8; the value of the log-likelihood evaluated at the maxi-

Figure 5.5: Plots of empirical and model predicted probabilities from 2-hit model with stochastic tissue growth: Age at first VS data (MUK data)

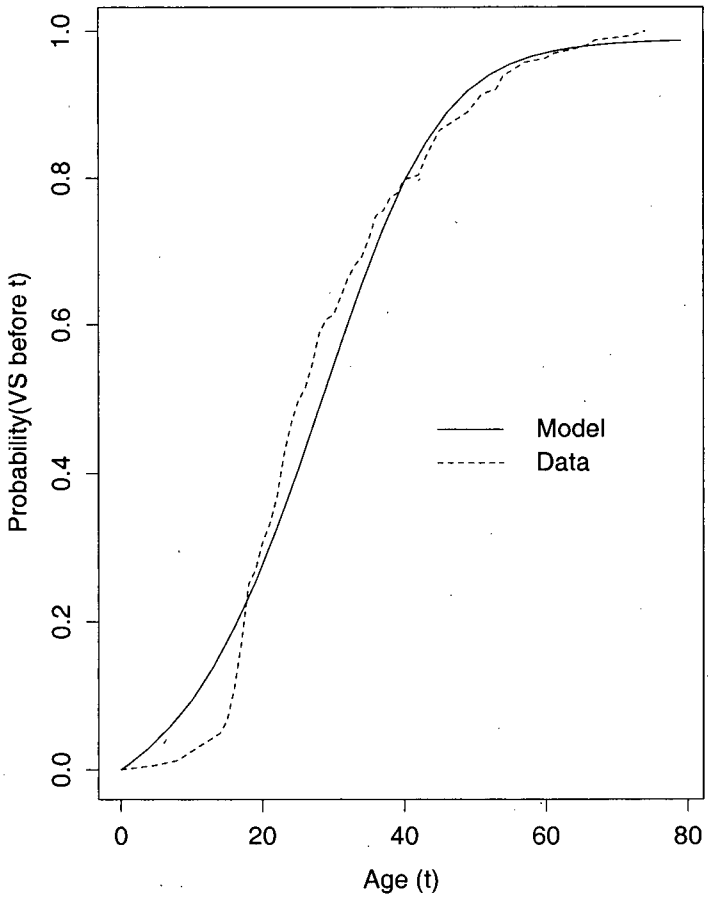


Figure 5.6: Plots of empirical and model predicted probabilities from 2-hit model with stochastic tissue growth: Age at hearing loss data (MUK data)

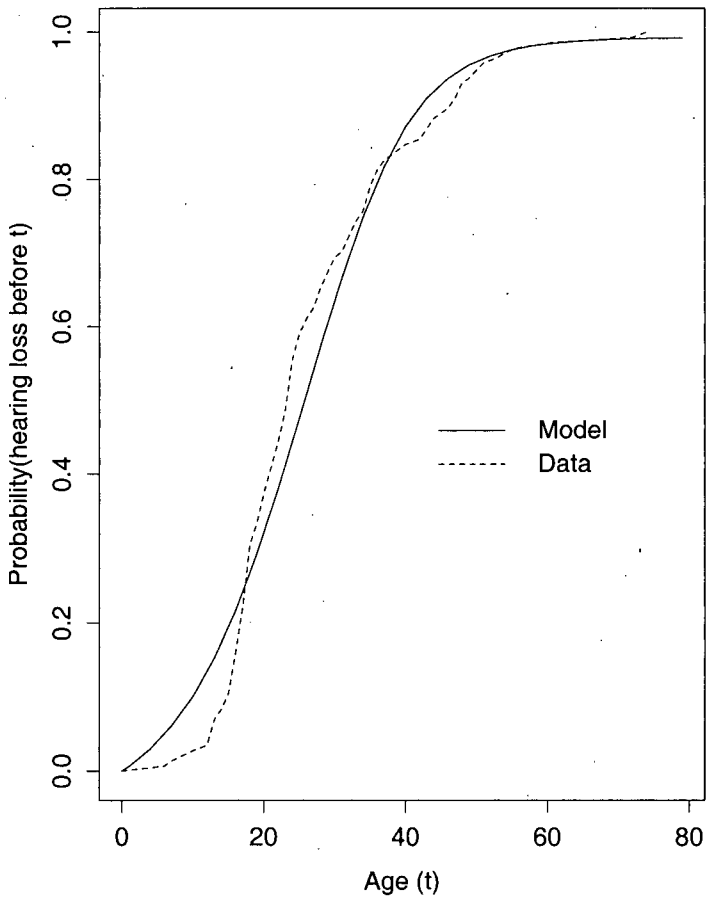


Figure 5.7: Plots of empirical and model predicted probabilities from 2-hit model with stochastic tissue growth: Age at hearing loss data (FSS data)

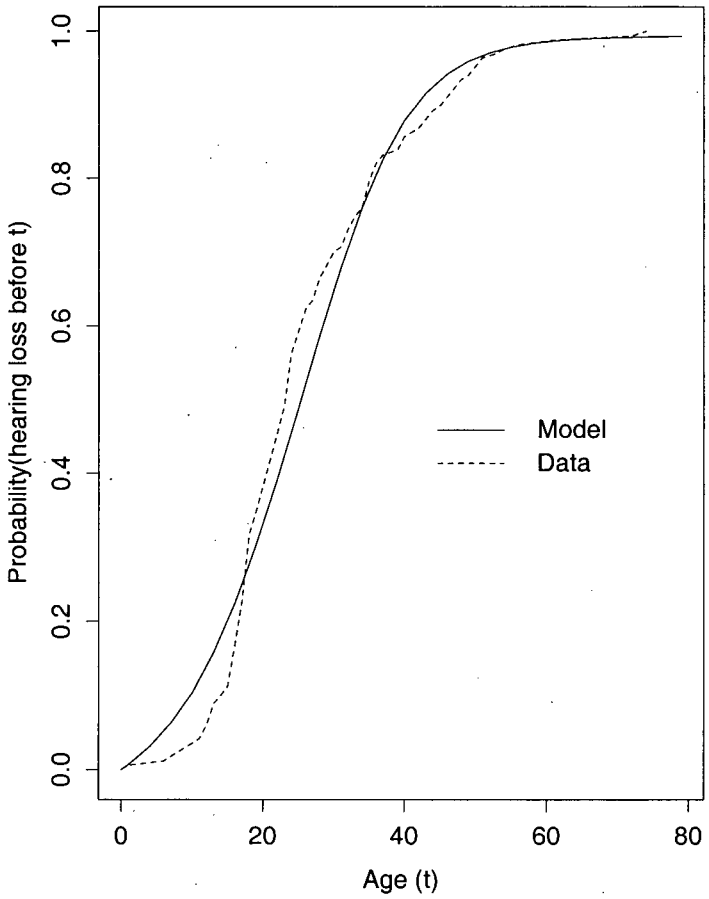
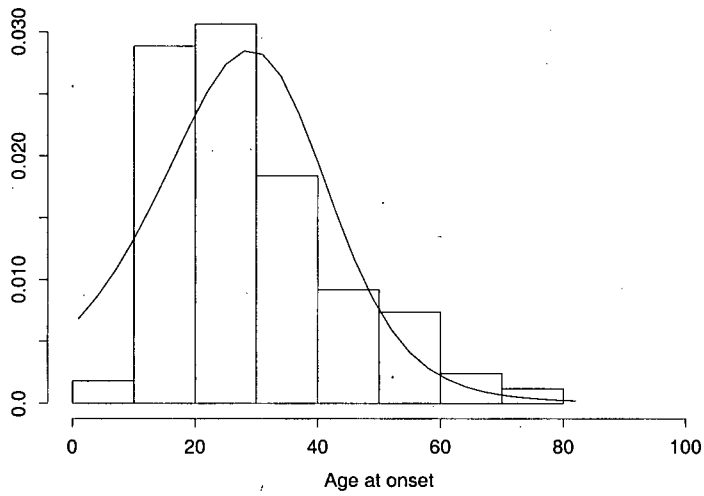
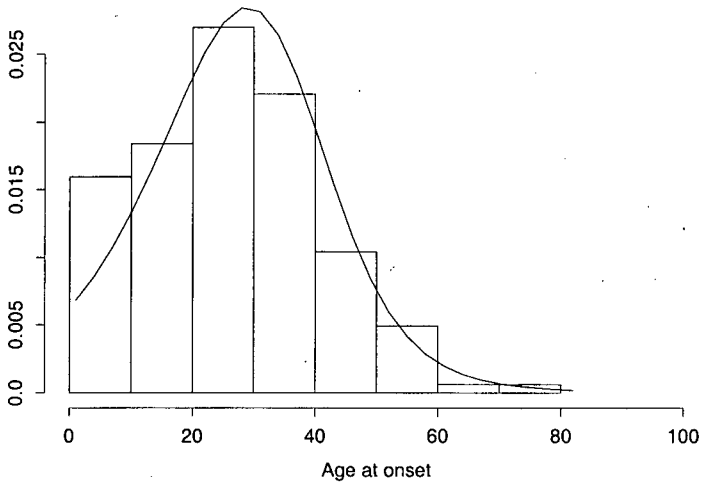


Figure 5.8: Histograms of data versus model simulated values for Age at first VS (MUK data); sample size of 163 patients

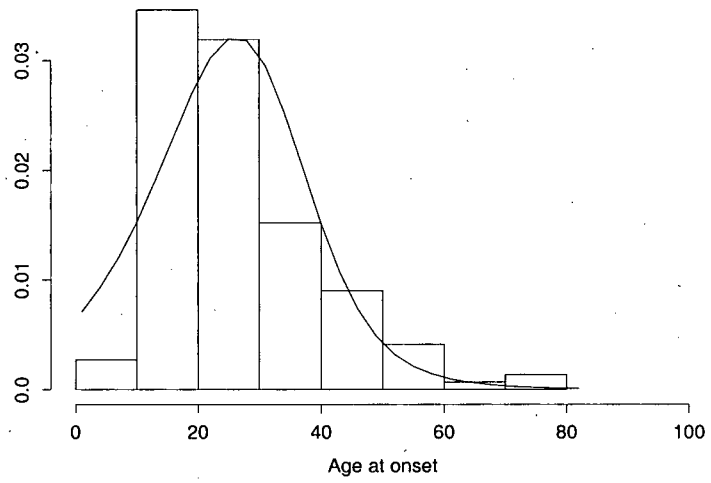


(a) Histogram of actual onset times

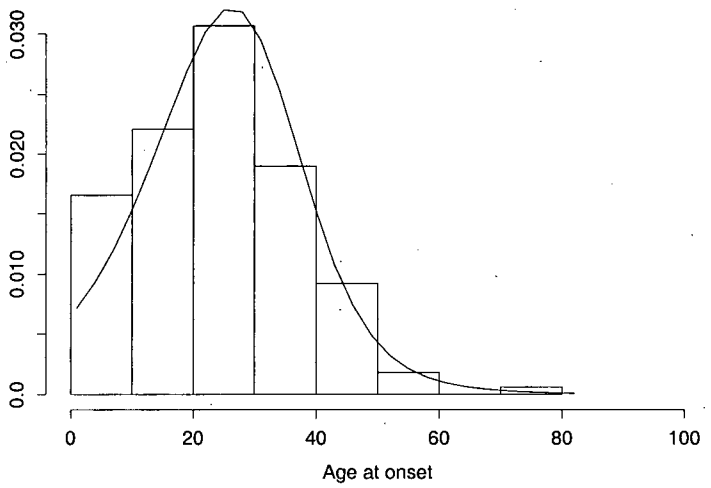


(b) Simulation from the fitted model

Figure 5.9: Histograms of data versus model simulated values for Age at Hearing loss (MUK data); sample size of 144 patients

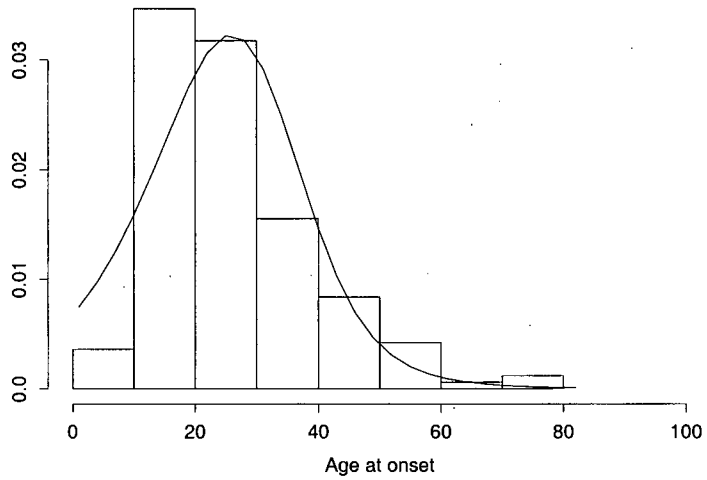


(a) Histogram of actual onset times

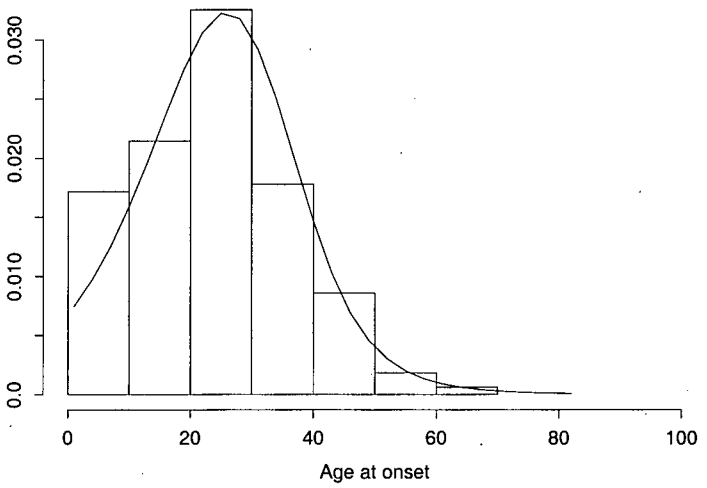


(b) Simulation from the fitted model

Figure 5.10: Histograms of data versus model simulated values for Age at Hearing loss (FSS data); sample size of 167 patients



(a) Histogram of actual onset times



(b) Simulation from the fitted model

Table 5.8: Model fitting results for 2-hit model incorporating genotype information (common mutation rates for both genotypes)

Variable: Age at onset of hearing loss		
Parameter	Estimate	Estimated SE
$\alpha^{(1)}$	10.231×10^{-1}	0.829×10^{-1}
$\alpha^{(2)}$	5.341×10^{-1}	1.887×10^{-1}
$\beta^{(1)}$	8.336×10^{-1}	0.816×10^{-1}
$\beta^{(2)}$	4.339×10^{-1}	1.788×10^{-1}
μ	2.272×10^{-4}	0.702×10^{-4}
value of log-likelihood: -253.747		

Table 5.9: Model fitting results for 2-hit model incorporating genotype information (different mutation rates for both genotypes)

Variable: Age at onset of hearing loss		
Parameter	Estimate	Estimated SE
$\alpha^{(1)}$	10.112×10^{-1}	3.109×10^{-1}
$\alpha^{(2)}$	5.203×10^{-1}	0.723×10^{-1}
$\beta^{(1)}$	8.248×10^{-1}	2.906×10^{-1}
$\beta^{(2)}$	4.208×10^{-1}	0.711×10^{-1}
$\mu^{(1)}$	2.397×10^{-4}	1.057×10^{-4}
$\mu^{(2)}$	2.318×10^{-4}	0.963×10^{-4}
value of log-likelihood: -253.755		

mum likelihood estimates is provided as well. Table 5.9 presents a similar summary of the parameter estimates and estimated standard errors from the six-parameter model. Note that the estimated model parameters are very similar for both models. In particular, the estimates for the mutation rates do not differ significantly across the different models. This supports our original intuition that the mutation rates for the two genotype groups should be equal. For both models, there are differences in in the estimates of the cell division and death rates across the two different genotype groups; the implications of this will be discussed below.

Figure 5.11 is a plot of the estimated and empirical cumulative distribution

functions for both genotypes; the estimated cumulative distribution functions have been computed using the fitted five-parameter model. The overall fit of the model is good; the most obvious concern would be that the model over-predicts the incidence of tumours for patients with genotype 2 for ages less than 15; this departure might be explained by the small sample sizes. Assessment of the fit is somewhat difficult given that the sample sizes from each of the genotype groups were quite small; recall that the sample sizes were 40 and 28 for genotype group 1 and genotype group 2 respectively. 95% confidence intervals for the empirical distribution function have been overlayed at ages 20 and 36 for both genotypes; these values were chosen only as examples. Figure 5.12 is an identical plot to that described above except that the estimated cumulative distribution functions have been computed using the fitted six-parameter model. This figure is almost indistinguishable from Figure 5.11; this is of course expected given the similarity in the parameter estimates given above. A likelihood ratio test confirmed that the addition of the extra mutation rate parameter did not result in a significantly better explanation of the data ($\chi^2_1 = 0.016$, p = 0.899).

There is a suggestion from the plots in Figures 5.11 and 5.12 that a patient's genotype affects the age at which they develop hearing loss (a surrogate for the age at onset of the first VS). Patients with protein truncating mutations clearly develop their hearing loss at an earlier age than patients with other mutation types. This observation is consistent with data from a previous study that observed NF2 patients with protein truncating mutations developing characteristic disease features at an earlier age than patients with spilce-site or missense mutations [7]. Although the six-parameter model allows the mutation rate parameters to differ across genotype groups, the estimates for the mutation rates are very similar for the two groups. This would suggest that the difference observed in the age at onset of hearing loss may be attributed to differences in the rates at which the pre-tumour cells divide and die across the two groups; this is a reasonable hypothesis to explain the differences

Figure 5.11: Plots of empirical and model predicted probabilities from genotype model with common mutation rates: Age at hearing loss data (FSS data)

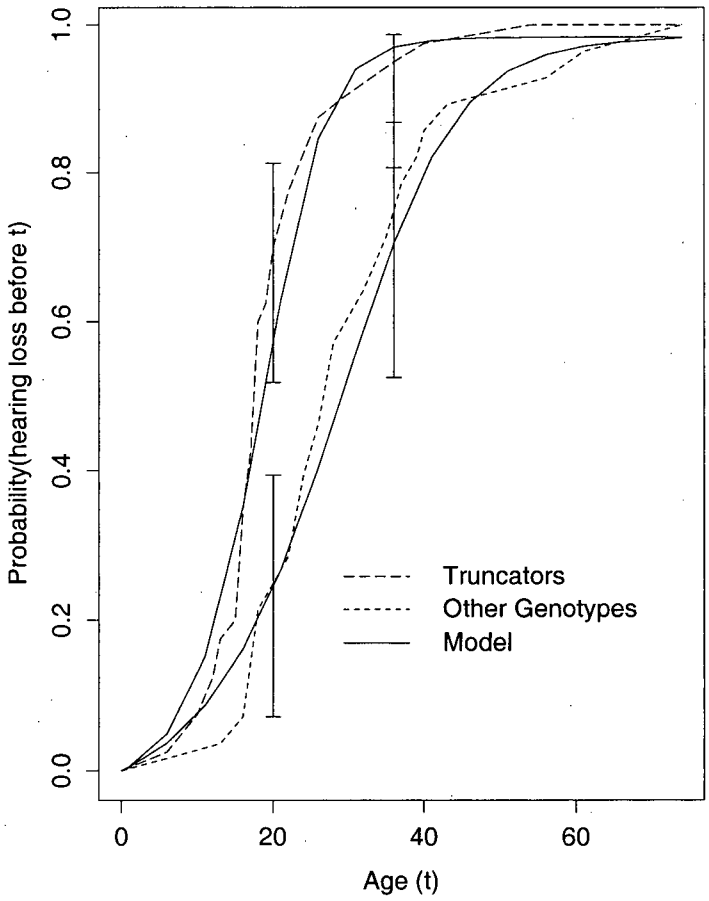
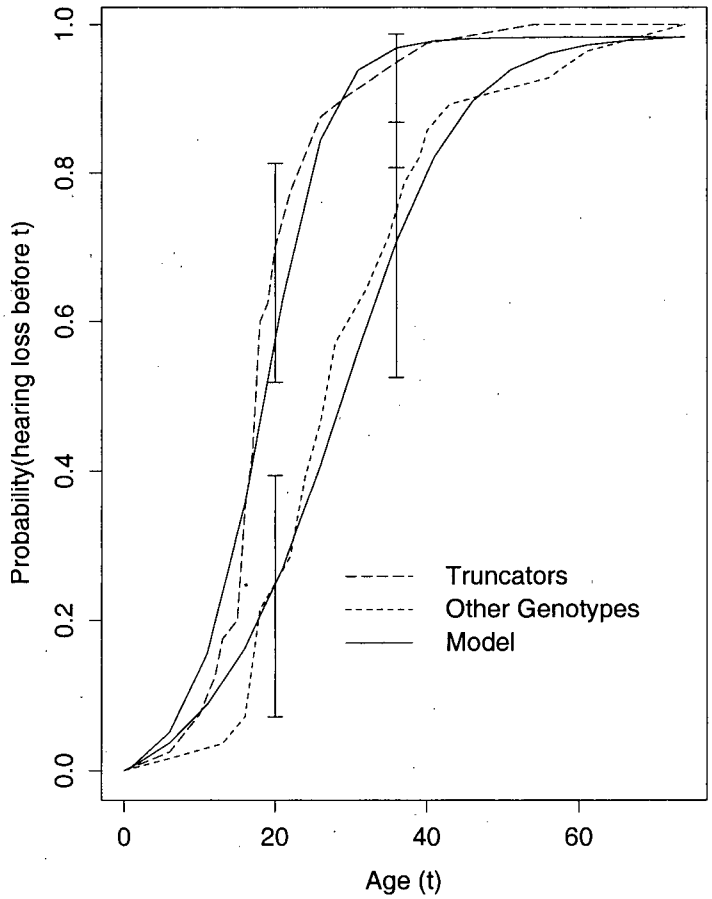


Figure 5.12: Plots of empirical and model predicted probabilities from genotype model with different mutation rates: Age at hearing loss data (FSS data)



in the age at onset of hearing loss across the two genotype groups.

A possible means of quantifying the differences in the cell division and death rates across genotypes might be the use of what Moolgavkar and Knudson [19] refer to as the 'growth advantage' of the pre-tumour cells. These authors explain that when the difference between the division rate and the death rate, $\alpha - \beta$, is positive, the pre-tumour cells are said to have a growth advantage; when this quantity is negative the pre-tumour cells are said to have a growth disadvantage. For genotype 1 the difference between the division and death rates is estimated as 1.895×10^{-1}, (based on the five-parameter model) indicating a growth advantage for the pre-tumour cells; this same estimate for genotype 2 is 1.002×10^{-1}, indicating a growth advantage as well. The growth advantage for genotype 1 is larger than it is for genotype 2 suggesting that the number of pre-tumour cells at risk of sustaining a chance mutation will likely be larger in patients with genotype 1 than in patients with genotype 2. This could perhaps explain the earlier age at onset of tumours in patients with protein truncating mutations. Moolgavkar and Knudson provide some interesting results related to the risk of tumour development in groups with different growth advantages using an approximate two-mutation model [19].

5.2.4 Estimating the Age at Onset of the Second VS

To fit the model for the age at onset of the second VS from Chapter 4 we must first specify the marginal distributions for the ages at onset of the left and right VSs. We stated previously that these two random variables are assumed to be identically distributed and discussed the justifications for this assumption. We have chosen to use the two-mutation model with stochastic growth of the tissue to specify the marginal distribution for the ages at onset of the left and right tumours. A discussion on the derivation of the hazard function for these random variables was provided in Chapter 4. We have only to specify the number of schwann cells initially present around a single vestibular nerve. We have assumed this quantity to be 10 cells; this is consistent with our previous assumption that there were 20 cells initially present around both the left and right nerves. The log-likelihood given in equation (4.2) was maximized with respect to the model parameters using a Quasi-Newton routine; this was implemented using programs written in the C programming language.

The model was fit to a subset of the MUK data; all 144 patients used for the fitting were probands with the ages at onset of both VSs recorded. Estimates of the four model parameters and their estimated standard errors are provided in Table 5.10; the value of the log-likelihood is also provided. The most striking estimate is the estimate of the dependence parameter δ. An estimate of this magnitude for δ corresponds to a value for Kendall's Tau of greater than 0.90 [10]; this indicates an extremely strong dependence between the ages at onset of the two tumours. Interpretating this parameter is somewhat delicate; it is possible that the dependence between the onset times for the two tumours that we have estimated here is artificial. For many individuals in our dataset, the ages at onset for the two tumours are recorded as the same age. The reason for this might be that many patients had already developed both tumours prior to their first examination. In this case the actual ages at onset are not really known and have been recorded as the first age that the patient was observed. The parameter estimate for δ is still meaningful here

perhaps, but it represents a different association. In this case it would represent the dependence between the ages at detection by physician of the left and right tumours; this may still be meaningful information for clinicians.

Table 5.10: Model fitting results for bilateral model using MUK data

Variables: Ages at onset of left and right VS		
Parameter	Estimate	Estimated SE
α	9.698×10^{-2}	3.191×10^{-2}
β	3.820×10^{-2}	2.542×10^{-2}
μ	8.041×10^{-4}	1.633×10^{-4}
δ	21.457	2.450
value of log-likelihood: -910.492		

Figure 5.13 are plots of the empirical distribution functions for the ages at onset of the first and second VS; Figure 5.14 is a similar display for the model fitted distribution functions. It is clear from Figure 5.13 that the ages at onset for the left and right tumours are very similar for the patients in our data. An inspection ot the data reveals that 122 of the 144 patients used in the fitting had both the left and right VSs present at the time of their first evaluation. There is slightly more separation between the model fitted distribution functions for the age at onset of the first and second tumours than is observed in the empirical distribution functions; the overall fit however, is quite reasonable. The distribution function for the age at onset of the second tumour estimated from the model is below the model fitted distribution function for the age at onset of the first tumour; this is the stochastic ordering that we would expect for these two distributions. Plots like the one shown in Figure 5.14 have an obvious value to physicians working with patients. Such plots would allow the clinician to explain the likely progression of disease for a patient with NF2 to the family of an affected individual.

Figure 5.13: Empirical distribution functions for the ages at onset of the first and second VS (MUK data)

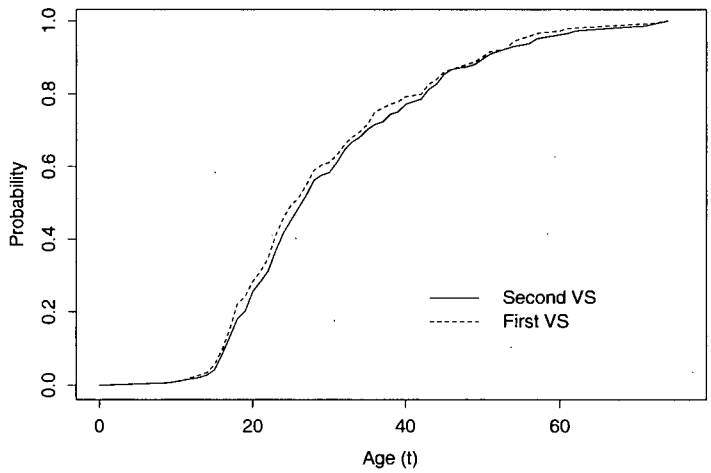
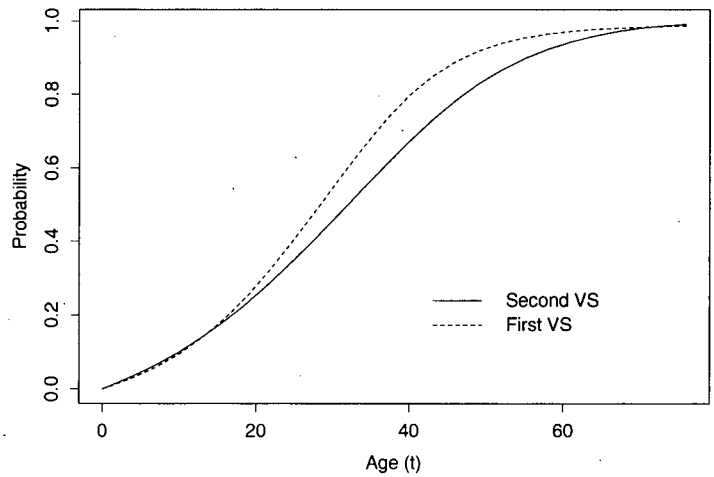


Figure 5.14: Model estimated distribution functions for the ages at onset of the first and second VS



5.2.5 3-hit Models

In order to fit the 3-hit model to our patient data we must first choose a deterministic function for the growth of the tissue at risk, similar to what was necessary for the deterministic 2-hit model. We will fit our models assuming two of the three growth functions previously used in the fitting of the deterministic 2-hit model; both the logistic$(10^7, 5.0, 0.8)$ and logistic$(10^7, 8.5, 1.3)$ will be employed. We will not display results from fitting the model assuming the third previously described growth function as our expectation is that they would be very similar to those obtained from assuming the logistic$(10^7, 8.5, 1.3)$ growth pattern. We have also chosen to display results from fitting the model to a single dataset, rather than all three datasets; we have chosen to use the age at onset of the first VS data for all 3-hit analyses. Again, it is our expectation that the results would be reasonably consistent across all three datasets. Results obtained from varying the first parameter in the growth functions from 10^7 to 10^6 will be presented below as well to demonstrate the sensitivity of the results to the choice of this parameter.

Tables 5.11 and 5.12 display the parameter estimates and the values of the log-likelihoods for the models fit assuming the logistic$(10^7, 5.0, 0.8)$ and logistic$(10^6, 5.0, 0.8)$ growth patterns respectively. Tables 5.13 and 5.14 are similar displays for the models fit assuming the logistic$(10^7, 8.5, 1.3)$ and logistic$(10^6, 8.5, 1.3)$ growth functions for the tissue. The values of the log-likelihoods for models fit assuming the logistic$(10^7, 5.0, 0.8)$ and logistic$(10^6, 5.0, 0.8)$ are essentially identical; this despite the fact that the parameter estimates differ quite a lot across these two models. This is also true for the models fit assuming the logistic$(10^7, 8.5, 1.3)$ and logistic$(10^6, 8.5, 1.3)$ growth functions, suggesting that the parameter estimates themselves are perhaps sensitive to the choice of the first parameter in the logistic growth functions; the value of the log-likelihood at the maximum likelihood estimate was not sensitive to the choice for this parameter value in our model fitting. This suggests that models fit assuming different growth functions could fit the data

Table 5.11: Model fitting results for 3-hit model using age at onset of the first VS (MUK Data); logistic(10^6,5,0.8) growth function

Variable: Age at onset of the first VS		
Growth Function: logistic(10^6,5,0.8)		
Parameter	Estimate	Estimated SE
α	2.991×10^{-1}	0.183×10^{-1}
β	3.175×10^{-1}	0.0929×10^{-1}
μ	5.594×10^{-5}	0.434×10^{-5}
value of log-likelihood: -643.300		

equally well, however some models may produce parameter estimates that are more interpretable than those obtained from other models. This is somewhat of a concern for us, as we are not certain about the number of cells that are present in an adult tissue.

Figures 5.15–5.18 are plots of the empirical and model estimated distribution functions from the fitting of the four aforementioned 3-hit models. The model fit is quite good for all four models; in particular the models fit assuming the logistic$(K, 8.5, 1.3)$, with $K = 10^6$ or 10^7, fit the data exceptionally well. The plots also support the claim cited above that models fit assuming different growth functions, for example growth functions that differ only in the first parameter, could fit the data equally well despite yielding different parameter estimates. The model estimated distribution functions from the fitting of the logistic$(10^7, 5.0, 0.8)$ and logistic$(10^6, 5.0, 0.8)$ are indistinguishable from one another; this can also be said for the model estimated distribution functions produced assuming the logistic$(10^7, 8.5, 1.3)$ and logistic$(10^6, 8.5, 1.3)$ growth functions.

Fitting any of the 3-hit models described above to patient data is less straightforward than the fitting of the 2-hit models described previously. Our experience has shown that the log-likelihoods for these models are quite flat and thus a very large sample is desired to facilitate the search for the maximum. The fitting was quite sensitive to the starting values used for the Quasi-Newton algorithm and con-

Table 5.12: Model fitting results for 3-hit model using age at onset of the first VS (MUK Data); logistic(10^7,5,0.8) growth function

Variable: Age at onset of the first VS		
Growth Function: logistic(10^7,5,0.8)		
Parameter	Estimate	Estimated SE
α	6.815×10^{-2}	0.00140×10^{-1}
β	7.003×10^{-1}	0.0467×10^{-1}
μ	1.773×10^{-5}	0.0615×10^{-5}
value of log-likelihood: -643.301		

Table 5.13: Model fitting results for 3-hit model using age at onset of the first VS (MUK Data); logistic(10^6,8.5,1.3) growth function

Variable: Age at onset of the first VS		
Growth Function: logistic(10^6,8.5,1.3)		
Parameter	Estimate	Estimated SE
α	3.825×10^{-1}	0.0433×10^{-1}
β	4.447×10^{-1}	0.294×10^{-1}
μ	7.575×10^{-5}	0.909×10^{-5}
value of log-likelihood: -639.988		

Table 5.14: Model fitting results for 3-hit model using age at onset of the first VS (MUK Data); logistic(10^7,8.5,1.3) growth function

Variable: Age at onset of the first VS		
Growth Function: logistic(10^7,8.5,1.3)		
Parameter	Estimate	Estimated SE
α	6.004×10^{-1}	0.947×10^{-1}
β	6.647×10^{-1}	0.744×10^{-1}
μ	2.417×10^{-5}	0.241×10^{-5}
value of log-likelihood: -639.990		

Figure 5.15: Plots of empirical and model predicted probabilities from 3-hit model assuming logistic(10^6,5,0.8) growth for the tissue; age at first VS data (MUK data)

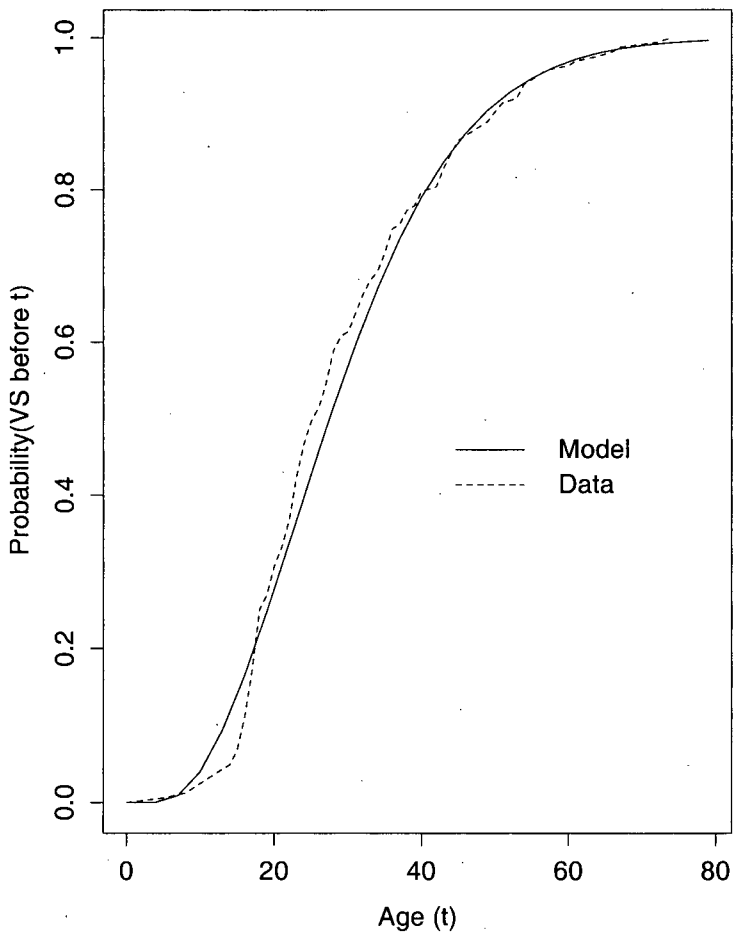


Figure 5.16: Plots of empirical and model predicted probabilities from 3-hit model assuming logistic(10^7,5,0.8) growth for the tissue; age at first VS data (MUK data)

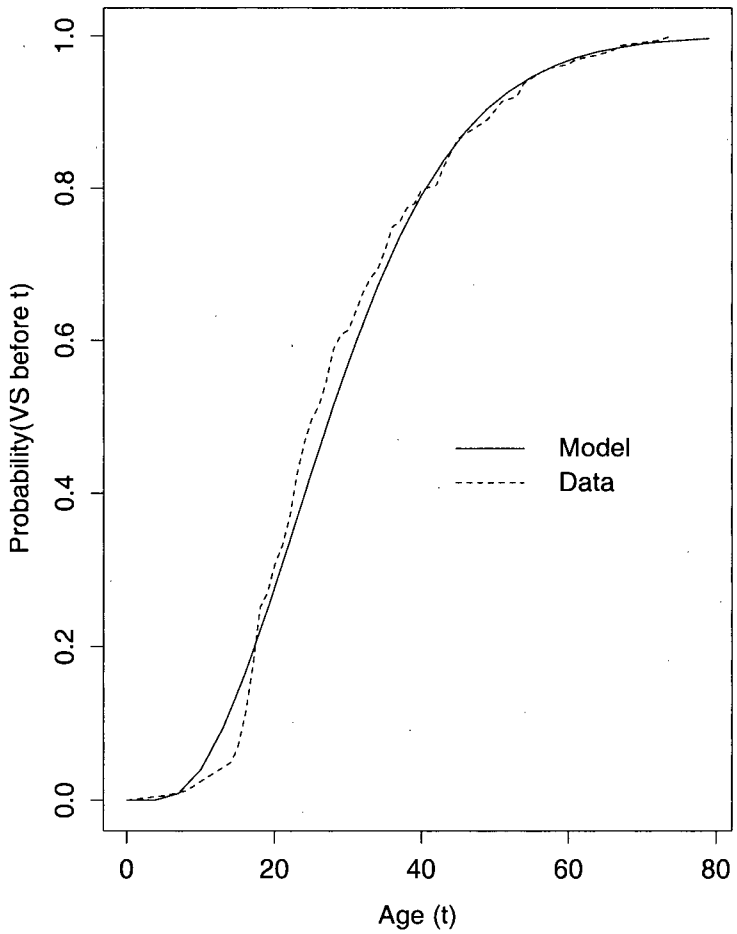


Figure 5.17: Plots of empirical and model predicted probabilities from 3-hit model assuming logistic(10^6,8.5,1.3) growth for the tissue; age at first VS data (MUK data)

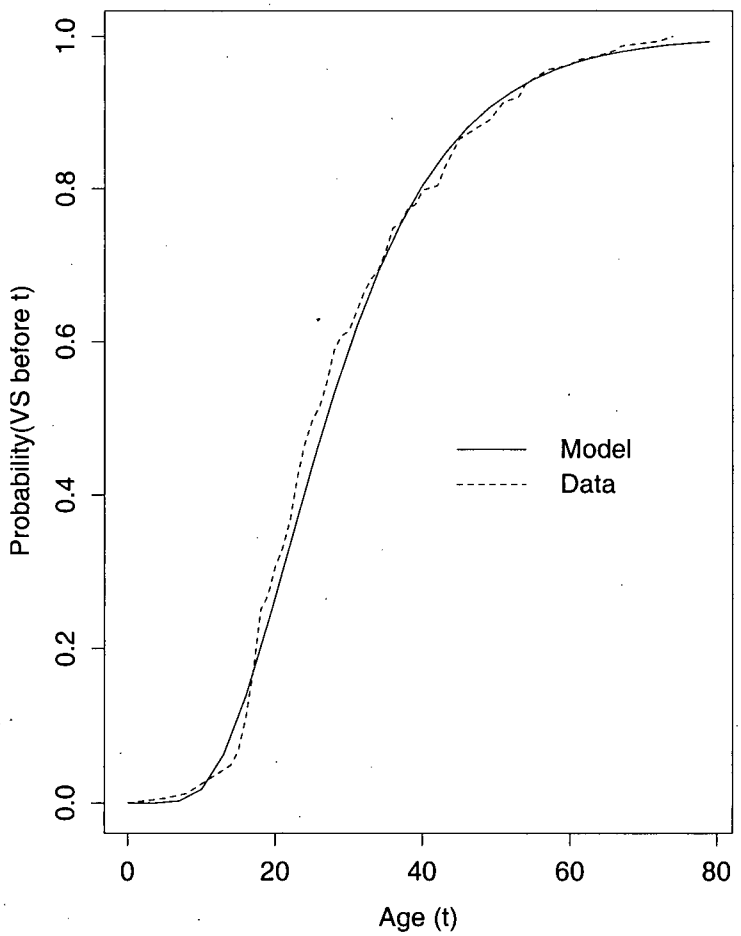
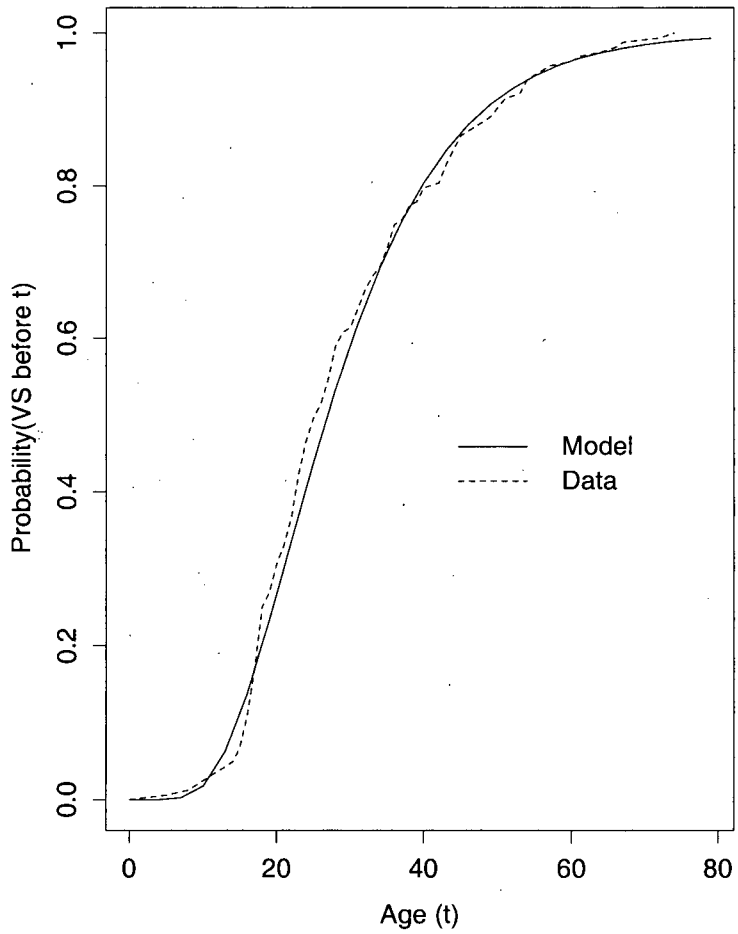


Figure 5.18: Plots of empirical and model predicted probabilities from 3-hit model assuming logistic(10^7,8.5,1.3) growth for the tissue; age at first VS data (MUK data)



vergence to a local maximum occurred for several starting values. Moolgavkar and Luebeck [21] reported that they found estimates of the growth and death rates to be unstable for such a model and suggested that greater stability might be obtained by re-parameterizing the model. We have not attempted to estimate model parameters under any alternative parameterizations at this point.

5.2.6 Comparison of Several Models

Several models have been presented in this thesis and results from the application of many of these models have appeared previously in this section. For a fixed dataset, it is natural to want to compare the fit of several models to the data to see if any one model provides a better fit than the others. In this section we will examine the fit of three of the models previously presented in this thesis, to the age at onset of the first VS variable from the MUK dataset; the three models will be the 2-hit model with deterministic tissue growth, 2-hit model with stochastic tissue growth and the 3-hit model. In particular, it is interesting to compare the fit of the 2-hit models with the 3-hit model, as this comparison is of biological importance. Such a comparison would allow us to examine which of the hypotheses for the development of tumour cells are most consistent with our data.

Results for the 2-hit model with deterministic growth of the tissue and for the 3-hit model will be presented here for the models that assume a logistic(10^7,8.5,1.3) growth function for the tissue. Table 5.15 is a summary of the log-likelihood values for the 3-hit model and for both the 2-hit model that assumes stochastic tissue growth and the 2-hit model that assumes deterministic tissue growth. The log-likelihood is largest for the 3-hit model and smallest for the 2-hit model that assumes deterministic growth of the tissue. The 2-hit model that assumes a stochastic growth of the tissue has a slightly larger log-likelihood value than the 2-hit model with deterministic tissue growth. One might have expected this to occur prior to the model fitting, as the fully stochastic model allows the parameters that govern the

Table 5.15: Comparison of log-likelihood values for different models; using age at onset of the first VS data (MUK data)

Variable: Age at onset of the first VS		
Model	log-likelihood	$\hat{\mu}$
2-hit (Deterministic tissue growth)*	-665.474	4.715×10^{-9}
2-hit (Stochastic tissue growth)	-659.966	3.154×10^{-4}
3-hit*	-639.990	2.417×10^{-5}

*Assuming a logistic(10^7,8.5,1.3) growth function for the tissue

growth and death of the tissue cells to be estimated from the data, whereas the deterministic model fixes these prior to the data analysis. It is interesting that the 3-hit model has the largest log-likelihood value of the three models. This indicates that this model provides a better fit to the data than either of the 2-hit models; this suggests that the 3-hit hypothesis might be more appropriate for the development of tumour cells than the 2-hit hypothesis (under the assumption that our models adequately represent these hypotheses). It is worth noting that the comparison of the 3-hit model with the 2-hit model that assumes deterministic growth of the tissue should be made using results that were obtained assuming a common growth function for the tissue across the two models.

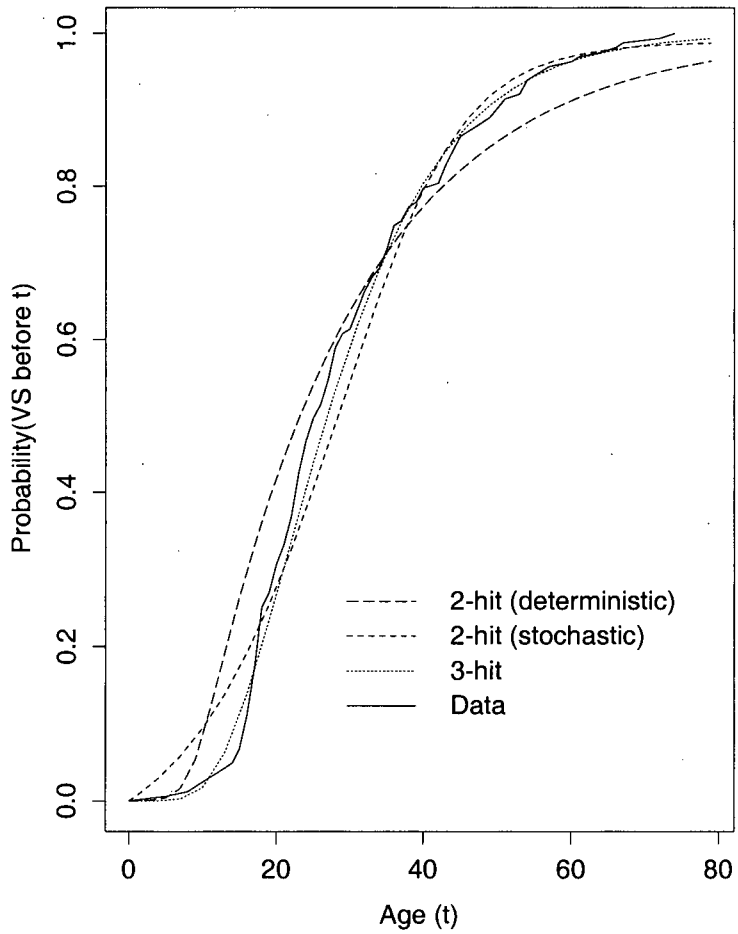
Figure 5.19 is a plot of the empirical distribution function for the data against the three model-estimated cumulative distribution functions. The estimated distribution function for the 3-hit model follows the empirical distribution function more closely than the distribution functions estimated from either of the 2-hit models. In particular, the 3-hit model fits very well for ages less than 20 years where both of the 2-hit models were observed to depart from the data. Again, there is suggestion from the plots of the estimated distribution functions that the 3-hit hypothesis for the development of tumour cells is most consistent with our data.

Although it is possible to compare the fit of several models to the data using log-likelihood values and plots, determining which hypothesis is the most

appropriate for the development of the tumour cells is still quite difficult. Our greatest concern is that the true growth function for the tissue is unknown to us and the results from both the 2-hit model with deterministic tissue growth, and the 3-hit model, depend heavily on the function chosen for the analysis. The acquisition of more information about the growth of the schwann cells would certainly improve our ability to compare the fit different models to the data.

An additional comparison to make might be to examine the magnitude of the estimated mutation rates from the three models described above. We have discussed previously that the mutation rate estimate for the 2-hit model with deterministic tissue growth depends directly on the value chosen for the number of cells in an adult tissue. The estimate for the mutation rate obtained from our fitting was 4.715×10^{-9}, which was considerably smaller than the estimate obtained from the model that assumes stochastic growth of the tissue (3.154×10^{-4}). We have not yet found a reference that suggests an appropriate value for this rate in patients with NF2 and thus comparing the rate estimates from the two models is difficult. Moolgavkar and Luebeck [21] reported mutation rate estimates in the range 3×10^{-8} to 4.5×10^{-7} from their analysis of colon cancer data using a 2-hit model; there is little reason however, to expect the rate estimates obtained from our fitting of the models to NF2 data to match their estimates. These same authors [21] also reported mutation rate estimates from an analysis of colon cancer data using a 3-hit model in the range 4.8×10^{-6} to 2.6×10^{-5}. The estimate of this rate from our analysis using the 3-hit model was 2.417×10^{-5}; again, there is little reason for us to expect a similarity in the estimates from our fitting and theirs. Discriminating between potential models for our data using the rate estimates as a criterion could only be done if information was available on the expected order of magnitude for these mutation rates under the different hypotheses for the development of the tumour cells.

Figure 5.19: Comparison of the estimated cumulative distribution functions for several models with the empirical distribution function for the age at onset of the first VS data (MUK data)



Chapter 6

Recommendations for Future Work

Future work could be surely improved with the addition of more precise information about the tissue at risk of developing the tumours. In particular, the number of schwann cells present in an adult tissue would be especially useful. We contacted several sources in an attempt to acquire such information and were unable to find the necessary information. We thank the individuals who took the time to reply to our queries related to this information. Knowledge of the periods at which the tissue undergoes spurts of growth would be useful as well. From our experiences, this information would not be as influential on the model fitting as would the number of cells in an adult tissue.

The most obvious suggestion for improving the fit of the models incorporating genotype information would be the acquisition of more patients with known mutation type. This would also allow us to make comparisons between more homogeneous genotype groups. The models fit in this thesis had to partition the patients into only two genotype groups; it would be interesting to explore these models with more than two genotype groups (e.g. splice-site mutations, protein-truncating mutations, missense mutations, other mutations).

The fitting of the model for the onset of the second VS could be improved with the acquisition of longitudinal data on NF2 patients. This would provide better information about the precise ages at onset of the left and right VS. Obviously, it is unrealistic to assume that this information could be found on probands. Probands are quite likely only going to be observed after being brought to medical attention for their condition. In our data, the presenting symptom for most patients was loss of hearing; this suggests that probands are monitored only after the onset of at least one VS making it very difficult to acquire accurate ages at onset for the two tumours. One suggestion here might be to use data on non-probands in the fitting; one non-proband per family could be randomly chosen to avoid issues of familial dependence. Non-probands, as a result of their family history, might be monitored more closely for the onset of the first and second VS. Our data did not contain enough non-probands with the necessary information to make the suggested fitting possible. In an unpublished manuscript [4], Evans *et al.* reported a mean difference between the ages at onset of the first and second VS of 5 years in non-probands; their study however, featured only 11 non-probands. It is still interesting to compare this mean difference in onset times to the mean difference of 1.05 years computed from our proband data. There is certainly a difference in sample sizes here that inhibits formal comparison, but still perhaps a suggestion that information on non-probands may be more informative for our modelling purposes.

The model fitting for the 3-hit model could perhaps be improved with the acquisition of more data. The likelihoods for the 3-hit models were quite flat and finding the global maximum was considerably more diffcult than it was with the 2-hit models. This problem might be reduced if the sample size was sufficiently larger. As well, reparameterizations of the 2-hit and 3-hit models could be explored to see if there is an improvement in numerical stability of the estimates and their estimated standard errors. Moolgavkar and Luebeck [21] suggest parameterizing both the 2-hit and 3-hit models with the mutation rate μ, the net cell division rate

$\alpha - \beta$, and the ratio of cell death to division β/α. These authors claim that this parameterization yields estimates that are more stable than those obtained under the parameterization presented in this thesis.

Bibliography

[1] Armitage, P., Doll, R., (1954). The age distribution of cancer and a multistage theory of carcinogenesis. *British Journal of Cancer,* **8**, 1-12.

[2] Chu, K.C., (1987). A nonmathematical view of mathematical models for cancer. *Journal of Chronic Diseases,* **40**,163S-170S.

[3] Clayton, D.G., (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika,* **65**, 141-151.

[4] Evans, D.G.R., Lye, R., Neary, W., Black, G, Strachan, T., Wallace, A., Ramsden, R.T. (date unknown). The probability of bilateral disease in individuals presenting with a unilateral vestibular schwannoma. *Unpublished Manuscript*

[5] Evans, D.G., Trueman, L., Wallace, A., Collins, S., Strachan, T. (1998) Genotype/phenotype correlations in type 2 neurofibromatosis (NF2): evidence for more severe disease associated with truncating mutations. *J Med Genet,* **35**, 450-455.

[6] Friedman, J.M., Gutmann, D.H., MacCollin, M., Riccardi, V.M., (1999). *Neurofibromatosis: Phenotype, Natural History, and Pathogenesis, 3rd ed.* Johns Hopkins University Press.

[7] Friedman, J.M., Woods, R., Joe, H., Evans, D.G.R., Wallace, A., Mautner, V.F., Kluwe, L., Parry, D.M., Rouleau, G.A., Baser, M.E. (1999). Germ-line

NF2 mutation type, location, and phenotype in neurofibromatosis 2. *American Journal of Human Genetics*, **65** (Supplement): pp. A126.

[8] Hethcote, H.W., Knudson, A.G.,(1978). Model for the incidence of embryonal cancers: application to retinoblastoma. *Proc. Nat. Acad. Sci.,* **75**, 2453-2457.

[9] Ince, E.L., (1926). *Ordinary Differential Equations.* London: Longmans, Green and Co. Ltd..

[10] Joe, H., (1997). *Multivariate Models and Dependence Concepts.* New York: Chapman & Hall, 139-149.

[11] Kimeldorf, G., Sampson, A.R., (1975). Uniform representations of bivariate distributions. *Comm. Statist.*, **4**, 617-627.

[12] Knudson, A.G.,(1971). Mutation and cancer: Statistical Study of retinoblastoma. *Proc. Nat. Acad. Sci.,* **68**, 820-823.

[13] Knudson, A.G., Hethcote, H.W., Brown, B.W., (1975). Mutation and childhood cancer: a probabilistic model for the incidence of retinoblastoma. *Proc. Nat. Acad. Sci.,* **72**, 5116-5120.

[14] Little, M.P., (1995). 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.

[15] Little, M.P., Muirhead, C.R., Stiller, C.A., (1996). Modelling lymphocytic leukemia incidence in England and Wales using generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon and Knudson. *Statistics in Medicine,* **15**, 1003-1022.

[16] Mathews, J.H., (1992). *Numerical Methods for Mathematics, Science, and Engineering, 2nd edition.* Englewood Cliffs, New Jersey: Prentice Hall.

[17] Mirz, F., Jorgensen, B., Fiirgaard, B., Lundorf, E., Pedersen, C.B., (1999). Investigations into the natural history of vestibular schwannomas. *Clinical Otolaryngology,* **24**, 13-18.

[18] Moolgavkar, S.H., Venzon, D.J., (1979). Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. *Mathematical Biosciences,* **47**, 55-77.

[19] Moolgavkar, S.H., Knudson, A.G., (1981). Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute,* **66**, 1037-1051.

[20] Moolgavkar, S.H., Dewanji, A., Venzon, D.J., (1988). A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Risk Analysis,* **8**, 383-392.

[21] Moolgavkar, S.H., Luebeck, G., (1990). Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. *Risk Analysis,* **10**, 323-341.

[22] Moolgavkar, S.H., Luebeck, G., (1992). Multistage carcinogenesis: population-based model for colon cancer. *Journal of the National Cancer Institute,* **84**, 610-617.

[23] Nash, J.C., (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation, 2nd edition.* New York: Hilger.

[24] Oakes, D., (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society, Series B,* **44** , 414-428.

[25] Parry, D.M., MacCollin, M.M., Kaiser-Kupfer, M.I., Pulaski, K., Nicholson, H.S., Bolesta, M., Eldridge, R., Gusella, J.F.,(1996). Germ-line mutations in the neurofibromatosis 2 gene: correlations with disease severity and retinal abnormalities. *Am J Hum Genet,* **59**, 529-539.

[26] Parzen, E., (1962). *Stochastic Processes.* San Francisco: Holden-Day.

[27] Ross, S.M., (1983). *Stochastic Processes.* New York: Wiley.

[28] Serio, G., (1984). Two-stage model for carcinogenesis with time-dependent parameters. *Statistics and Probability Letters,* **2**, 95-103.

Appendix A

General Solution of the Riccati Equation (equation (3.6))

Equation (3.5) is a partial differential equation of the form:

$$P\frac{\partial \phi}{\partial t} + Q\frac{\partial \phi}{\partial x} = R,$$

which has general solution of the form $g(u, v) = 0$, where

$$u(t, x, \phi) = a \text{ is the solution to } Pdx = Qdt$$

$$u(t, x, \phi) = b \text{ is the solution to } Pd\phi = Rdt.$$

In our problem $P = 1$, $R = 0$ and $Q(x, z) = -[\mu xz + \alpha x^2 + \beta - (\alpha + \beta + \mu)x]$.

Let s be the antiderivative of $1/Q$ with respect to x. Then $Pdx = Q(x)dt$ becomes $dx = Q(x)dt$ or $s(x) = t + a$. $Pd\phi = Rdt$ becomes $d\phi = 0$ or $\phi = b$. Thus, the general solution is of the form $g(s(x) - t, \phi) = 0$. Applying the boundary condition $\phi(x, z, 0) = x$ at $t = 0$ we get $g(s(x), x) = 0$ or $g(u, v) = u - s(v)$ or

$$s(x) - t - s(\phi) = 0.$$

Differentiating this last expression with respect to t yields:

$$-1 = s'(\phi)\frac{\partial \phi}{\partial t} = [Q(\phi)]^{-1}\frac{\partial \phi}{\partial t}.$$

Rearranging this leads directly to $\frac{\partial \phi}{\partial t} = -Q(\phi)$ which is the Riccati equation (equation (3.6)).

Additional Justification for Equation (3.11)

Here we justify the expression for $\Psi'(y,z;t)$ used in the derivation of the three-hit hazard function. Recall that:

$$\begin{aligned}
\Psi'(y,z;t) &= \frac{\partial \Psi(y,z;t)}{\partial t} \\
&= \sum_{j,k=0}^{\infty} p'((j,k);t) y^j z^k \\
&= \sum_{j,k=0}^{\infty} [-j(\alpha+\beta+\mu)\cdot p((j,k);t) - \mu X(t)\cdot p((j,k);t) \\
&\quad +(j-1)\alpha \cdot p((j-1,k);t) + (j+1)\beta \cdot p((j+1,k);t) \\
&\quad +\mu X(t)\cdot p((j-1,k);t) + j\mu \cdot p((j,k-1);t)\,]\, y^j z^k \\
&= \sum_{j,k=0}^{\infty} [-j(\alpha+\beta+\mu)\cdot p((j,k);t)]\, y^j z^k - \sum_{j,k=0}^{\infty} \mu X(t)\cdot p((j,k);t) y^j z^k \\
&\quad + \sum_{j,k=0}^{\infty} (j-1)\alpha \cdot p((j-1,k);t) y^j z^k \\
&\quad + \sum_{j,k=0}^{\infty} (j+1)\beta \cdot p((j+1,k);t) y^j z^k \\
&\quad + \sum_{j,k=0}^{\infty} \mu X(t)\cdot p((j-1,k);t) y^j z^k \\
&\quad + \sum_{j,k=0}^{\infty} j\mu \cdot p((j,k-1);t) y^j z^k \qquad \text{(A.1)}
\end{aligned}$$

Now we examine each term from the previous expression in more detail:

$$\begin{aligned}
\sum_{j,k=0}^{\infty} [-j(\alpha+\beta+\mu)\cdot p((j,k);t)]\, y^j z^k &= -(\alpha+\beta+\mu) y \sum_{j,k=0}^{\infty} j p((j,k);t) y^{j-1} z^k \\
&= -(\alpha+\beta+\mu) y \cdot \frac{\partial \Psi(y,z;t)}{\partial y}
\end{aligned}$$

$$\sum_{j,k=0}^{\infty} \mu X(t)\cdot p((j,k);t) y^j z^k = \mu X(t) \Psi(y,z;t)$$

$$\sum_{j,k=0}^{\infty}(j-1)\alpha\cdot p((j-1,k);t)y^j z^k = y^2\alpha\sum_{j,k=0}^{\infty}(j-1)p((j-1,k);t)y^{j-2}z^k$$
$$= y^2\alpha\cdot\frac{\partial\Psi(y,z;t)}{\partial y}$$

$$\sum_{j,k=0}^{\infty}(j+1)\beta\cdot p((j+1,k);t)y^j z^k = \beta\sum_{j,k=0}^{\infty}(j+1)p((j+1,k);t)y^j z^k$$
$$= \beta\cdot\frac{\partial\Psi(y,z;t)}{\partial y}$$

$$\sum_{j,k=0}^{\infty}\mu X(t)\cdot p((j-1,k);t)y^j z^k = y\mu X(t)\sum_{j,k=0}^{\infty}p((j-1,k);t)y^{j-1}z^k$$
$$= y\mu X(t)\Psi(y,z;t)$$

$$\sum_{j,k=0}^{\infty}j\mu\cdot p((j,k-1);t)y^j z^k = \mu yz\sum_{j,k=0}^{\infty}jp((j,k-1);t)y^{j-1}z^{k-1}$$
$$= \mu yz\cdot\frac{\partial\Psi(y,z;t)}{\partial y}$$

Substituting these 6 expressions above into equation (A.1) we obtain the following:

$$\Psi'(y,z;t) = (y-1)\mu X(t)\Psi(y,z;t) + [\mu yz + \alpha y^2 + \beta - (\alpha+\beta+\mu)y]\frac{\partial\Psi(y,z;t)}{\partial y}$$

which is the expression given previously in the derivation of the three-hit hazard function.

Appendix B

Glossary

allele alternative forms of a genetic locus; a single allele for each is inherited separately from each parent.

autosome a chromosome not involved in sex determination. The human genome consists of 22 pairs of autosomes and the sex chromosomes.

dominant allele the allele in a heterozygous state that determines the phenotype.

exon the coding region of a gene.

gene location on chromosome p=short arm, q=long arm; the location of a gene is often given by specifying the chromosome number and the letter representing the arm of the chromosome on which it is found.

genotype the entire genetic constitution of an organism. Our definition will more specifically refer to alleles at a given locus.

intron the area between coding regions of the gene [i.e. the regions between exons]

locus position of a gene on a chromosome.

meningioma a tumour on the coverings of the brain and spinal cord.

protein truncating mutation examples of such mutations include nonsense [premature stop codon from base-pair change], frameshift [bases involved not a multiple of three] and deletions; these mutations affect the DNA coding at a locus in that protein is not fully produced.

splice site mutation a mutation between the exon and intron.

missense mutation base-pair change such that there is a substitution of one amino acid for another.

phenotype any genetically controlled, observable property of an organism.

proband the first family member to be brought to medical attention for their condition or genetic disease.

retinoblastoma a malignant tumour of the eye that arises in the retinal cells, usually occurring in children, with a frequency of 1 in 20000. Associated with a deletion on the long arm of chromosome 13 (13q).

tinnitus a ringing of the ears.

vestibular schwannoma a tumour of the schwann cells that line the vestibular nerves.