SURVIVAL ANALYSIS ON PEDIGREES: A MARKED POINT PROCESS MODEL

BY

ANGUS S. MACDONALD

ABSTRACT

Regulation of insurers' use of genetic information means actuaries are interested in age-at-onset of genetic disorders. Arjas & Haara (1984) suggested marked point processes (MPPs) as useful models for life history data with complex covariates. Age-at-onset distributions (or equivalently, hazard rates) in respect of inherited disorders are often estimated from pedigrees, which are life histories with unusually complex covariates, as well as strong dependencies induced by shared genes. Since Elston (1973) parametric models have often been used, conditioning the likelihood on known genotypes. However, a genotype identified by a presymptomatic genetic test is a form of internal covariate (Kalbfleisch & Prentice, 2002). We propose a very general MPP model of a pedigree, including presymptomatic genetic testing, ('the full model') and show under what circumstances the partial model leading to Elston's likelihood is valid. In practice, pedigrees are often ascertained retrospectively. Many such events can be modelled by augmenting the natural filtration of the MPP. We show that, except in simple special cases, the partial model is no longer valid, and the resulting likelihoods appear to be intractable. In particular, ascertainment interacts even with independent censoring so that likelihoods no longer factorize. For one simple special case — studies of sibships — we generalise a classical result to age-at-onset data. We conclude that the study of genetic conditions with variable age at onset gains insights from the underlying principles of survival analysis in their modern form, but that great care is needed in translating epidemiological studies into actuarial models.

KEYWORDS

Ascertainment; Insurance; Internal Covariate; Marked Point Process; Pedigree; Presymptomatic Genetic Test; Rate of Onset; Survival Analysis.

1. INTRODUCTION

1.1. Survival Analysis on Pedigrees

Medical underwriting of long-term insurance contracts often must be based on the epidemiological literature. The typical study seeks to estimate onset rates

Astin Bulletin 40(1), 35-64. doi: 10.2143/AST.40.1.2049218 © 2010 by Astin Bulletin. All rights reserved.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

of disease, and mortality rates after onset. The actuary's premium rates, based on these estimates, are then themselves point estimates, which must inherit sampling distributions depending on the data and the study design. This aspect of medical underwriting is rarely considered, in contrast to its non-life equivalent.

Genetics has forced the pace, because the furore over insurers possibly using presymptomatic genetic test results has led to this kind of information being regulated. In particular, in the U.K., the government has charged the Genetics and Insurance Committee (GAIC) with deciding when a genetic test result is so relevant to the insurance risk that it may be used in underwriting (for large risks only). The criteria adopted by GAIC, although not explicitly statistical, suggest that its decisions might be aided by understanding the sampling properties of premium rate estimates. We would expect GAIC to require, as a minimum, that premium rate estimates should be unbiased, and be reliably distinct for different genotypes.

Lu, Macdonald & Waters (2008) and Lu, Macdonald & Wekwete (2008) studied the dispersion of critical illness (CI) insurance premium rate estimates. They used rates of disease onset derived from survival analyses of family histories, assuming these rates to be unbiased estimates. However, Espinosa & Macdonald (2007) showed that one of the usual assumptions made in survival analysis — that truly independent censoring does not introduce bias — does not hold if the pedigrees were sampled retrospectively, as is often the case. They considered non-parametric (Nelson-Aalen) estimation, whereas many pedigree studies use parametric (likelihood) models. Our motivation here, with actuarial applications in mind, is the question of bias in parametric survival analysis based on pedigrees.

To see why problems might arise, contrast the features of a 'normal' survival study — prospective design, independent subjects, random censoring and fixed covariates — with a typical pedigree study. Following GAIC, we focus on the presence or absence of a rare heritable gene variant that leads to disease onset in adults, and note that disease onset is the event of interest, and death a form of censoring.

(a) Rarity prevents sampling of the general population, so subjects are drawn from families in which the disease has been reported.
(b) At least one family member must be affected for the family to be noticed by researchers. The latter is an event, not always well defined, called 'ascertainment'. Inference should not be based on the information that led to ascertainment. Failure to exclude it leads to 'ascertainment bias'. There is a large literature on ascertainment bias and how to adjust for it (see Hodge (2002) and references therein) but little in the specific context of onset rates and survival analysis. Li (2007) is a useful reference for survival analysis in genetic epidemiology.
(c) Because family members share genes, their lifetimes are dependent. (They will share environment too, but that is beyond our scope.) Likelihoods must be summed over all possible joint genotypes for the entire pedigree.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

After allowing for any genotypes that have been observed directly, Mendel's laws and the family structure give the probability distribution for the joint pedigree genotype.

(d) The advent of DNA sequencing means that a person's genotype may be revealed by a healthy person taking a *presymptomatic* genetic test. The usual reason for doing so is that they have one or more affected relatives, so they are members of families more likely to be ascertained anyway. The result of any such test is an *internal covariate*. That is, we learn the test result *because* the person concerned remained free of onset up to the time of the test. This is discussed in Section 1.2.

(e) Late onset means that observation of onset may be censored in the usual way.

Arjas & Haara (1984) proposed marked point process (MPP) models for survival data with complicated covariates. Pedigrees with presymptomatic genetic testing fit that description, and have highly structured dependencies between lifetimes also. In any survival analysis, we hope that all complicating factors (of which the simplest is censoring) can be ignored, in the sense that while they may affect the form of the likelihood, we need only estimate those parameters related directly to the event of interest. This often amounts to using a partial model instead of a full model that includes all complicating factors. However, only by specifying the full model can we tell whether or not the use of a partial model is justified (Arjas & Haara, 1984). Of particular interest to us is whether the method of ascertainment may interact with censoring, so that the latter cannot be ignored.

Elston (1973) defined pedigree likelihoods for genetic traits with late onset, allowing rates of onset to be estimated (see Section 1.4). This work pre-dated genetic testing, hence known genotypes as (internal) covariates. It also preceded the modern point process approach to survival analysis, which is the key to testing whether the full model leads to a valid partial model.

Our aim is to formulate a MPP model of a pedigree, to address the following questions.

(a) Can we recover Elston's likelihood from a properly formulated survival model?

(b) How are internal covariates generated by presymptomatic genetic tests to be included in the model?

(c) How does retrospective ascertainment: (i) interact with censoring; and (ii) affect estimates of onset rates?

1.2. Genetic Test Results Are Internal Covariates

Suppose a man had a presymptomatic genetic test at age 40 and was shown to carry a disease-causing mutation. In estimating the rate of onset at age 30 (say) in a survival analysis we ask: "given that he is healthy at age 30, what is the probability that onset occurs before age 31?". Knowing the genotype, a

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

sharper question is: "given that he is healthy at age 30 and carries a causative mutation, what is the probability that onset occurs before age 31?". But, allowing for how we learned about the genotype, we should really ask: "given that he is healthy at age 30 and a causative mutation was revealed by a presymptomatic test at age 40, what is the probability that onset occurs before age 31?", to which the answer is "zero".

This is the problem of an 'internal' covariate (Kalbfleisch & Prentice, 2002) whose value is learned by observing an event which is: (a) of interest only because it reveals the covariate; but (b) not independent of the event we are studying. Worse, in a pedigree, a presymptomatic test carried out on one person will change what we know about other people's genotypes, because of Mendel's laws.

A discrete-time example (based on Arjas & Haara (1984, Section 3)) shows how to finesse internal covariates. Denote the 'events' in a life history A_1, A_2, …, A_n in order of their occurrence. They may be events of interest, 'nuisance' events like censoring, or time passing while nothing happens. Suppose a covariate takes values in a set $\mathcal{U}$. If the observed covariate is $v \in \mathcal{U}$, and the model parameter is θ, the likelihood constructed in the usual sequential way is:

$$\begin{aligned} L(\theta;n) &= P[A_1, A_2, \ldots, A_n, v] \\ &= P[v]\, P[A_1 \,|\, v]\, P[A_2 \,|\, v, A_1] \cdots P[A_n \,|\, v, A_1, \ldots, A_{n-1}] \end{aligned} \tag{1}$$

from which $P[v]$ may be dropped if the covariate distribution does not depend on θ. But how did we learn v? Strictly, we did so by observing an event C, which must be part of the model. Sometimes C is trivial, for example, if everyone is genotyped at the start of a prospective study. But if C happens at a random time it carries information about some of the A_i. Suppose that A_i is the event 'free of symptoms at age i' and C is the event 'had a presymptomatic genetic test at random age j with result v'. Since the test precedes onset by definition, $C \Rightarrow A_i$ for $i \le j$, and the sequential construction is:

$$\begin{aligned} L(\theta;n) &= P[A_1]\, P[A_2 \,|\, A_1] \cdots P[A_j \,|\, A_1, \ldots, A_{j-1}] \times P[C | A_1, \ldots, A_j] \\ &\quad \times P[A_{j+1} \,|\, C]\, P[A_{j+2} \,|\, C, A_{j+1}] \cdots P[A_n \,|\, C, A_{j+1}, \ldots, A_{n-1}]. \end{aligned} \tag{2}$$

It seems that we should condition on v only after it is known. However, the event C has two parts: the random age j, and the result v. Denote the first of these C_j, so $C = (C_j, v)$. In MPP terms, v is a mark observed when C_j occurs. The likelihoods just before and after the genetic test are:

$$L(\theta; j^-) = P[A_1]\, P[A_2 \,|\, A_1]\, P[A_3 \,|\, A_1, A_2] \cdots P[A_j \,|\, A_1, \ldots, A_{j-1}] \tag{3}$$

$$= \sum_{u \in \mathcal{U}} P[u]\, P[A_1 \,|\, u]\, P[A_2 \,|\, u, A_1]\, P[A_3 \,|\, u, A_1, A_2] \cdots \tag{4}$$
$$P[A_j \,|\, u, A_1, \ldots, A_{j-1}]$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

$$L(\theta; j^+) = L(\theta; j^-)\, P[C_j \mid A_1, \ldots, A_j]\, P[v \mid C_j, A_1, \ldots, A_j]. \tag{5}$$

From Bayes' Theorem:

$$P[v \mid A_1, \ldots, A_j] = \frac{P[v]\, P[A_1 \mid v]\, P[A_2 \mid v, A_1] \cdots P[A_j \mid v, A_1, \ldots, A_{j-1}]}{\sum_{u \in \mathcal{U}} P[u]\, P[A_1 \mid u]\, P[A_2 \mid u, A_1] \cdots P[A_j \mid u, A_1, \ldots, A_{j-1}]} \tag{6}$$

$$= \frac{P[v]\, P[A_1 \mid v]\, P[A_2 \mid v, A_1] \cdots P[A_j \mid v, A_1, \ldots, A_{j-1}]}{L(\theta; j^-)} \tag{7}$$

so, if $P[C_j \mid A_1, \ldots, A_j]$ does not depend on θ, and if $P[v \mid C_j, A_1, \ldots, A_j] = P[v \mid A_1, \ldots, A_j]$ (so the test result depends only on the same history that may have caused the test to be taken at age j), we have:

$$L(\theta; j^+) \propto P[v]\, P[A_1 \mid v]\, P[A_2 \mid v, A_1] \cdots P[A_j \mid v, A_1, \ldots, A_{j-1}] \tag{8}$$

as if the genotype v was known at outset. Further, for $k > j$ assume that $P[A_k \mid C, A_{j+1}, \ldots, A_{k-1}] = P[A_k \mid v, A_1, \ldots, A_{k-1}]$; then we recover the likelihood (1). In terms of Arjas & Haara (1984) these assumptions mean that the only innovative part of the genetic test was the result v. Arjas & Haara pointed out that only by investigating the full model (including the event $C = (C_j, v)$ as above) can we tell whether or not we may use a partially specified model, such as one that just conditions on v. However, the example above does not immediately suggest how to introduce internal covariates when we have dependent life histories linked by shared genes.

1.3. The Elements of a Genetic Model of Inherited Disease

Rates of onset, the focus of actuarial interest, are just one part of a genetic model of inherited disease. We assume the whole model has a parameter denoted θ.

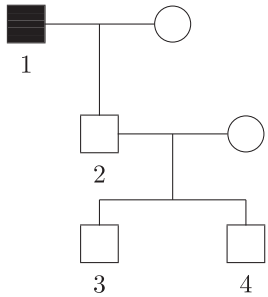


FIGURE 1: A three-generation pedigree in respect of a dominantly inherited disorder. Squares are males, circles are females, and a filled-in symbol denotes onset of the disease.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

(a) The frequencies with which different versions of a gene (called alleles) occur are parts of the model.
(b) The physical characteristics conferred by an allele, called the phenotype, depend on how it acts on processes in the body. If one deleterious allele is sufficient to cause a disease, the disease is dominantly inherited, and can be passed on by just one parent. If two deleterious alleles are necessary, the disease is recessively inherited, and must be passed on by both parents.
(c) Some diseases are purely genetic in origin, for example Huntington's disease has no known cause except mutations in the huntingtin gene. But many common diseases, such as breast or colorectal cancer, have inherited variants.
(d) Even the presence of a proven 'causative' genotype does not guarantee that disease will be observed. The probability of onset by age x, given the genotype and assuming all other decrements to be absent, is called the age-related penetrance of the genotype. Penetrance at high ages may be less than 100%, and in practice other decrements introduce censoring.

1.4. Pedigree Likelihoods

Let θ be the parameter of a genetic model, defining rates of onset and mode of transmission. Let there be M genotypes, labelled from 1 to M, and N pedigree members, labelled from 1 to N. The possible joint genotypes of the pedigree are the N-tuples in the set $\mathcal{U} = \{1, 2, \ldots, M\}^N$. The probability, under θ, that the pedigree genotype is u is denoted $P_\theta[u]$. Let $\mu_g(x, \theta)$ be the hazard of onset at age x, given genotype g. The hazard is interpreted as follows: the probability that a healthy person of genotype g and age x will suffer onset by age $x + dx$ is $\mu_g(x, \theta)\, dx$, for small dx. Suppose the highest age at which the jth person was observed is x_j^*, and give the indicator $\mathbf{I}_j$ the value 1 if x_j^* was the age at onset or the value 0 if x_j^* was the age at censoring (we assume that observation ceases at onset). Then from Elston (1973) the likelihood is:

$$L_1(\theta) = \sum_{u \in \mathcal{U}} \prod_{j=1}^{j=N} \left[\exp\left(-\int_0^{x_j^*} \mu_{u_j}(x, \theta)\, dx \right) \mu_{u_j}(x_j^*, \theta)^{\mathbf{I}_j} \right] P_\theta[u]. \tag{9}$$

(Elston (1973) parameterized the likelihood in terms of the lifetime penetrance, and the distribution of age at onset conditional on onset occurring. In (9), penetrance is incomplete (less than 100%) if $\int_0^\infty \mu_g(x, \theta)\, dx < \infty$, defining an improper distribution for the age at onset. Here, it will be clearer if the hazard is explicit.)

As an example, consider the pedigree in Figure 1, in respect of a rare dominantly inherited disorder such that: (a) there are only two genotypes, 'normal' labelled 1 and 'mutation' labelled 2; and (b) the mutation is the sole cause of the disease. Let ϕ be the population prevalence of mutation carriers. The pedigree genotype is $u = (u_1, u_2, u_3, u_4)$, the indices referring to the labels

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

TABLE 1

TERMS IN THE LIKELIHOOD IN RESPECT OF THE PEDIGREE IN FIGURE 1, BEFORE AND AFTER THE PERSON LABELLED '4' SUFFERS ONSET AT AGE x_4. WILD-TYPE ALLELES ARE LABELLED 1, AND MUTATIONS ARE LABELLED 2.

Genotype	Before Onset	After Onset
(2,1,1,1)	$\phi\ 0.5\, f_2(x_1)\ S_1(x_2)\ S_1(x_3)\ S_1(x_4)$	$\phi\ 0.5\, f_2(x_1)\ S_1(x_2)\ S_1(x_3)\, f_1(x_4) = 0$
(2,2,1,1)	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_1(x_3)\ S_1(x_4)$	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_1(x_3)\, f_1(x_4) = 0$
(2,2,2,1)	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_2(x_3)\ S_1(x_4)$	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_2(x_3)\, f_1(x_4) = 0$
(2,2,1,2)	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_1(x_3)\ S_2(x_4)$	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_1(x_3)\, f_2(x_4) \neq 0$
(2,2,2,2)	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_2(x_3)\ S_2(x_4)$	$\phi\ 0.125\, f_2(x_1)\ S_2(x_2)\ S_2(x_3)\, f_2(x_4) \neq 0$

in the figure. Table 1 shows the contributions to the likelihood (9) just before and just after person 4 suffers onset at age x_4. Because the mutation is rare, we ignore the possibility that it is introduced to the pedigree by more than one person, so (for example) pedigree genotype $(2,1,2,1)$ is considered infeasible. For brevity we drop θ and define the survivor functions and densities, respectively as:

$$S_g(x) = \exp\left(-\int_0^x \mu_g(y)\,dy\right) \quad \text{and} \quad f_g(x) = S_g(x)\,\mu_g(x). \tag{10}$$

Implicit in Elston's likelihood is the assumption that pedigree members' life histories are independent, conditional on the pedigree genotype. Then the likelihoods before and after the onset at age x_4 are found by summing the columns in Table 1; for example the likelihood just after the onset at age x_4 is proportional to:

$$\begin{aligned} &P[u=(2,2,1,2)]\ f_2(x_1)S_2(x_2)S_1(x_3)f_2(x_4) \\ &\qquad + P[u=(2,2,2,2)]\ f_2(x_1)S_2(x_2)S_2(x_3)f_2(x_4)\,. \end{aligned} \tag{11}$$

Impossible genotypes (sufferers cannot have genotype 1) vanish whenever onset occurs because $\mu_1(x,\theta) = 0$ for all θ, which is not the same as conditioning upon known genotypes, although it looks like it.

If, instead of suffering onset, person 4 had taken a genetic test at age x_4, we would hope to proceed by defining $\mathcal{U}^* \subset \mathcal{U}$ to be the set of pedigree genotypes consistent with known genotypes *as revealed by genetic tests*, then using a likelihood proportional to:

$$L_2(\theta) = \sum_{u\in\mathcal{U}^*} \prod_{j=1}^{j=N} \left[\exp\left(-\int_0^{x_j^*} \mu_{u_j}(x,\theta)\,dx\right) \mu_{u_j}(x_j^*,\theta)^{\mathbf{I}_j}\right] P_\theta[u]. \tag{12}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

However, test results are internal covariates and, as in Section 1.2, we cannot just sum over $\mathcal{U}^*$ instead of $\mathcal{U}$, without including genetic testing in a fully specified model. Moreover, unlike most covariates in a survival model, a genetic test result can ripple through the whole pedigree, for example altering what was previously known about deceased ancestors. Our approach, in Sections 3 and 4, is to define $\mathcal{U}^*(t)$ to be the set of pedigree genotypes consistent with all genetic test results known at time t. We show that a new member joining at time t expands $\mathcal{U}^*(t^-)$ (just before time t) to a larger set $\mathcal{U}^*(t) \supset \mathcal{U}^*(t^-)$; that a genetic test taken at time t shrinks $\mathcal{U}^*(t^-)$ to a smaller set $\mathcal{U}^*(t) \subset \mathcal{U}^*(t^-)$; and that the likelihood based on the history at time t is the sum (12) over $\mathcal{U}^*(t)$.

2. Survival Analysis Based on Marked Point Processes

2.1. Marked Point Processes and Hazards

As references for this section, Andersen *et al.* (1993) is a standard text on survival analysis based on counting processes, while Arjas & Haara (1984) and Arjas (1989) introduce the MPP approach.

The idea of a MPP is that events happen at a sequence of random times $t_1, t_2, \ldots$, and at each time t_r information called a 'mark' may be obtained, for now denoted e_r. Marks may convey almost any kind of information, three important examples being the following:

(a) The mark e_r may indicate that a specific type of event has occurred, for example a transition between two distinct states of health. Thus a multistate or staging model is a MPP.
(b) The mark e_r may indicate that an event happened to a particular person. That is, we can model the joint life histories of several people as a single MPP. This is merely cosmetic if they have mutually independent life histories, but not in the case of pedigree data.
(c) The mark e_r may be the value of a covariate observed at time t_r, for example, the result of a presymptomatic genetic test

We work in continuous time, and suppose that: (a) we observe persons, labelled $i = 1, 2, \ldots$, whose life histories need not be mutually independent; and (b) any one person's life history is modelled by transitions between a finite set of states $\mathcal{S}$. Then an event h is defined by the triple (i, k, l), meaning that the ith person made a transition between states k and l of $\mathcal{S}$. Let $\mathcal{H}$ be the set of all such possible events.

Event h can happen at time t only if the ith person is in state k just before time t, which time we denote t^-. Then we say the MPP is 'at risk of event h at time t^-'. Define the following processes:

$$\mathbf{Y}_h(t) = \begin{cases} 1 & \text{if at risk of event } h \text{ at time } t^- \\ 0 & \text{if not at risk of event } h \text{ at time } t^- \end{cases} \tag{13}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

$$\mathbf{N}_h(t) = \text{the number of occurrences of event } h \text{ by time } t \tag{14}$$

$$\lambda_h(t) = \text{the hazard rate of event } h \text{ at time } t \tag{15}$$

$$\mathbf{A}_h(t) = \int_0^t \mathbf{Y}_h(s)\,\lambda_h(s)\,ds. \tag{16}$$

Let $\mathcal{F}_t$ ($\mathcal{F}_{t^-}$) denote the information obtained by observing the MPP up to and including (not including) time t. Interpret hazard rates as in Section 1.4, but allowing for $\mathcal{F}_{t^-}$:

$$P[\text{Event } h \text{ occurs before time } t + dt \mid \mathcal{F}_{t^-} \text{ and at risk of event } h \text{ at time } t^-] \approx \lambda_h(t)\,dt. \tag{17}$$

To refer to events $h = (i,k,l)$, we will use the compact notation $\lambda_h(t) = \lambda^i_{kl}(t)$ instead of $\lambda_{(i,k,l)}(t)$ (and similarly for $\mathbf{Y}_h(t)$, $\mathbf{N}_h(t)$ and $\mathbf{A}_h(t)$).

2.2. Likelihoods

Heuristically, given the history at time t^-, the event h happens at time t with probability $\mathbf{Y}_h(t)\,\lambda_h(t)\,dt = d\mathbf{A}_h(t)$, or does not happen with probability $1 - d\mathbf{A}_h(t)$. These give the Bernoulli likelihood:

$$\big(1 - d\mathbf{A}_h(t)\big)^{1-d\mathbf{N}_h(t)}\big(d\mathbf{A}_h(t)\big)^{d\mathbf{N}_h(t)}. \tag{18}$$

Multiplying all these terms over the interval $[0,T]$, in the limit, the likelihood is the product integral:

$$\prod_{[0,T]} \big(1 - d\mathbf{A}_h(t)\big)^{1-d\mathbf{N}_h(t)}\big(d\mathbf{A}_h(t)\big)^{d\mathbf{N}_h(t)}. \tag{19}$$

The simplest example is $h = (i,1,2)$, where state 1 is 'alive' and state 2 is 'dead', and supposing the ith person was observed until time t^*; (19) is:

$$\begin{cases} \exp\left(-\int_0^{t^*} \lambda^i_{12}(s)\,ds\right)\lambda^i_{12}(t^*) & \text{if observation ends with death at time } t^* \\ \exp\left(-\int_0^{t^*} \lambda^i_{12}(s)\,ds\right) & \text{if observation is censored at time } t^*. \end{cases} \tag{20}$$

Each event $h \in \mathcal{H}$ contributes a factor like (19) to the likelihood, so:

$$L(\theta,T) = \prod_{h\in\mathcal{H}} \prod_{[0,T]} \big(1 - d\mathbf{A}_h(t)\big)^{1-d\mathbf{N}_h(t)}\big(d\mathbf{A}_h(t)\big)^{d\mathbf{N}_h(t)}. \tag{21}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

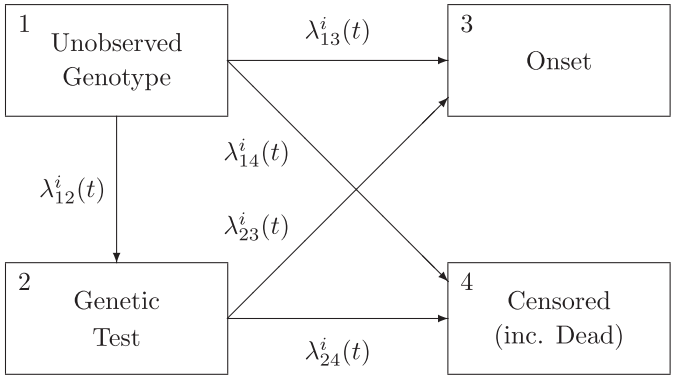


FIGURE 2: A model of the life history of the ith member of a pedigree, in State 1 from birth, with genotype being then unobserved. Each possible transition from State j to State k is governed by a hazard rate (intensity) $\lambda^i_{jk}(t)$ depending on calendar time t.

2.3. Marks

Information, called a 'mark', may be acquired when an event $h \in \mathcal{H}$ occurs at time t. It is denoted $e_h(t)$, taking values in a suitable mark space, denoted $E_h(t)$. Let $\mathcal{F}_t$ now include the information obtained by observing marks. Then the likelihood (21) becomes:

$$L(\theta, T) = \prod_{h\in\mathcal{H}} \prod_{[0,T]} (1 - d\mathbf{A}_h(t))^{1-d\mathbf{N}_h(t)} \left(d\mathbf{A}_h(t)\, P\left[e_h(t) \,\middle|\, \mathcal{F}_{t-},\, d\mathbf{N}_h(t) = 1\right]\right)^{d\mathbf{N}_h(t)}. \tag{22}$$

For brevity, we will usually write $P[e_h(t)]$ instead of $P[e_h(t)\,|\,\mathcal{F}_{t-}, d\mathbf{N}_h(t) = 1]$, the conditioning on the event occurring being understood.

3. A Marked Point Process Model of a Pedigree

3.1. Aims

We model a pedigree as a MPP with the following observable events: new members joining, disease onset, predictive (presymptomatic) genetic tests and censoring. We leave aside obvious extensions, for example to diagnostic testing after onset. We need a lot of notation, so we list the main definitions at the end of the paper.

3.2. Timescales

We need to track all pedigree members simultaneously in calendar time, so we index all processes by calendar time t, only introducing age x as needed (see Andersen *et al.* (1993), Section X.1.9). Observation takes place in the

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

interval $[0, T]$ of calendar time. Suppose the ith person is born at calendar time b_i; then define:

$$x_i(t) = t - b_i = \text{the age of the } i\text{th person at calendar time } t.$$

3.3. The Genetic Model: Genotypes and Penetrances

Let there be M genotypes, labelled $1, 2, \ldots, M$, and let the population frequency of genotype g be ϕ_g. For genotype g define the hazard or intensity of onset, a function of age x, denoted as $\mu_g(x)$:

$$\mu_g(x) = \text{Rate of onset at age } x \text{ given genotype } g. \tag{23}$$

Thus, if the ith pedigree member has (or is assumed to have) genotype g, the rate of onset to which they are subject at calendar time t, if then at risk, is $\mu_g(x_i(t))$. Knowing $\mu_g(x)$ is equivalent to knowing the penetrance of genotype g. Denote the parameter of the genetic model θ (so ϕ_g and $\mu_g(x)$ are part of θ), and denote the resulting probability distribution P_θ.

3.4. Life Histories

The ith pedigree member's life history is represented by the states and transitions shown in Figure 2, that is, the set of events:

$$\mathcal{H}_i = \{(i,1,2), (i,1,3), (i,1,4), (i,2,3), (i,2,4)\} \tag{24}$$

and the set of events in the pedigree model is $\mathcal{H} = \bigcup_i \mathcal{H}_i$. Note that the hazards (intensities) $\lambda_h(t) = \lambda^i_{kl}(t)$ are functions of calendar time, and they are not in general identical for different pedigree members (so the 'i' superscript in $\lambda^i_{kl}(t)$ is not superfluous).

(a) 'Unobserved genotype' (state 1) indicates that the ith member has not had a genetic test or suffered onset, so all knowledge about his/her genotype probabilities comes from his/her relatives.
(b) State 4 accounts for all censoring, except observation ceasing at time T. We treat death as a form of censoring.
(c) A genetic test at time t is accompanied by a mark, the genotype. The mark space is $E^i_{12}(t) = \{1, 2, \ldots, M\}$.
(d) The intensities $\lambda^i_{kl}(t)$ are conditioned on what is known at time t^-, so $\lambda^i_{13}(t) \neq \lambda^i_{23}(t)$.

3.5. The Pedigree History

We need to add structure to express the dependencies in a pedigree. To do this, define a new event, not represented in Figure 2, namely new members joining.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

This is the essential difference between modelling the pedigree and modelling the trivial collection of its members. The idea is as follows:

(a) At calendar time t, the pedigree has $N(t)$ members (alive or dead; having joined, no-one leaves).
(b) Events happen at a sequence of random times $t_1, t_2, \ldots$.
(c) Events may be of two kinds: (i) one of the events $h \in \mathcal{H}$ defined above (an existing pedigree member experiences a transition); or (ii) one or more new members join the pedigree, either by birth or by marriage. More than one new member may join only in the case of a multiple birth.

Augment $\mathcal{H}$ with a new event labelled '0', which is 'one or more persons join the pedigree': define $\mathcal{H}^* = \{0\} \cup \mathcal{H}$. Joiners may be 'founders', meaning original members or spouses of members, or 'births', meaning children born to existing members. Let $\mathbf{N}_0(t)$ be the number of joining events by time t (note that $\mathbf{N}_0(t) \le N(t)$) and let $\lambda_0(t)$ be the associated intensity. With each joining event, associate a mark $e_0(t)$ that carries all the information that may be needed. For births, this must identify the parents, and siblings in the case of multiple births. For marriages, it should include sex, the age $t - b_i$ on joining the pedigree, and which state in Figure 2 was entered upon joining. The intensity $\lambda_0(t)$ will be a function of the numbers and ages of all potential childbearing couples in the pedigree, and possibly extra information known at time t^- as well. Any likelihood from which $\lambda_0(t)$ cannot be dropped will almost certainly be intractable, but we need to specify a full model before seeing what can be omitted in a partial model.

For simplicity, we suppose that joiners enter State 1 in Figure 2. This is realistic in the case of births, and can easily be relaxed for spouses joining the pedigree.

3.6. Pedigree Genotypes

We invent a 'null' genotype labelled 0 to which members not yet in the pedigree are assigned.

(a) Let $\mathcal{U}(t)$ be the set of all possible pedigree genotypes at time t. That is, the set of all $u = (u_1, u_2, \ldots, u_{N(t)}, 0, 0, \ldots)$, where each $u_i \in \{1, 2, \ldots, M\}$ for $i \le N(t)$.
(b) Also, let $\mathcal{U}^*(t)$ be the set of all possible pedigree genotypes allowing for any members whose genotype is known *because of a genetic test* taken not later than time t. For example, if the 1st pedigree member has been tested before time t and has genotype v_1, then $\mathcal{U}^*(t) \subseteq \{u \in \mathcal{U}(t) : u_1 = v_1\}$.

3.7. Observation and Information

(a) The information known at time t from observing the pedigree history is denoted $\mathcal{F}_t$.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

(b) The information known at time t solely from knowing the pedigree structure (that part of the 'joining' marks $e_0(t)$ identifying parents and identical siblings) is denoted $\mathcal{A}_t$.

3.8. Genotype Probabilities

We define two sets of genotype probabilities: those conditioned on the pedigree structure alone ($\mathcal{A}_t$) and those conditioned on all observations ($\mathcal{F}_t$).

(a) Define the genotype probability at time t conditional on the pedigree structure at time t to be:

$$p_u(t) = P_\theta[u \mid \mathcal{A}_t] \quad (u \in \mathcal{U}(t)). \tag{25}$$

To define genotype transmission probabilities, suppose the ith member joins at time t, and let u' be elements of $\mathcal{U}(t^-)$. The transmission probability, denoted p_u^i, is as follows:

$$p_u^i = P_\theta[u_i \mid u', e_0(t)] \tag{26}$$

where $u = (u'_1, \ldots, u'_{i-1}, u_i, 0, 0, \ldots)$. If the ith member is a founder, $p_u^i = \phi_{u_i}$. Otherwise, if $e_0(t)$ identifies the fth member as his/her father and the mth member as his/her mother, p_u^i is the probability that a father with genotype u_f and a mother with genotype u_m have a child with genotype u_i.

(b) Define the pedigree genotype probability at time t conditional on all that is known then to be:

$$m_u(t) = P_\theta[u \mid \mathcal{F}_t] \quad (u \in \mathcal{U}(t)). \tag{27}$$

The corresponding probability that the ith member has genotype $g \in \{1, 2, \ldots, M\}$ we define as:

$$m_g^i(t) = \sum_{\substack{u \in \mathcal{U}(t) \\ u_i = g}} m_u(t). \tag{28}$$

3.9. Rates of Onset in Calendar Time

It is clear that $\lambda_{23}^i(t) = \mu_g(x_i(t))$, where g is the known genotype, and it is easily shown that the $\lambda_{13}^i(t)$ are averages of the $\mu_g(x_i(t))$, weighted by the genotype probabilities just defined:

$$\lambda_{13}^i(t) = \sum_{v_i=1}^{v_i=M} m_{v_i}^i(t^-)\, \mu_{v_i}(x_i(t)). \tag{29}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

3.10. Likelihoods

Our problem can now be stated. Elston's likelihood (9) is defined in terms of age-related onset rates and genotype probabilities as in Section 3.8(a) above, conditioned on $\mathcal{A}_T$, the ultimate pedigree structure at time T. Also, it does not allow for genetic testing as an internal covariate. The likelihood obtained from the survival model, set out in (30) below, is stated in terms of intensities in calendar time, and genotype probabilities must be conditioned on $\mathcal{F}_t$, all history up to time t, as in Section 3.8(b) above. We need to show that they are equivalent.

4. Elston's Likelihood

4.1. Definitions and Notation

The likelihood, in terms of the calendar-time intensities $\lambda^i_{jk}(t)$, is:

$$L(\theta;T) = \prod_{h\in\mathcal{H}^*}\prod_{[0,T]}(1-d\mathbf{A}_h(s))^{1-d\mathbf{N}_h(s)}\big(d\mathbf{A}_h(s)P[e_h(s)]\big)^{d\mathbf{N}_h(s)}. \tag{30}$$

assuming random sampling of the first member. We now show that this is proportional to the likelihood (12), extending Elston's likelihood (9) allowing for genetic tests.

In this section, we suppose that $\lambda^i_{13}(t)$ and $\lambda^i_{23}(t)$ are the only intensities depending on genotype or θ, and drop other intensities from the likelihood whenever they appear as factors. Define:

$$\mathcal{R}_t = \{i : \mathbf{Y}^i_{13}(t) = 1 \text{ or } \mathbf{Y}^i_{23}(t) = 1\}$$

to indicate all those at risk of onset just before time t.

We will often use sums over all feasible pedigree genotypes, such that the ith person's genotype v_i is fixed, $\sum_{u\in\mathcal{U}^*(t):u_i=v_i}$, so for brevity let $\sum_{u_i=v_i}$ represent this sum.

4.2. Sequential Construction of $L(\theta;T)$

We will construct $L(\theta;T)$ sequentially, conditioning the probabilities of events in each time increment on the past. Heuristically:

$$L(\theta;t) = L(\theta;t^-)\times P_\theta[\text{Events observed at time } t \mid \mathcal{F}_{t^-}]. \tag{31}$$

The key point is that:

$$L(\theta;t^-) = P_\theta[\mathcal{F}_{t^-}] = \sum_{u\in\mathcal{U}^*(t^-)} P_\theta[\mathcal{F}_{t^-},u] \tag{32}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

so:

$$m_v(t^-) = P_\theta[v \mid \mathcal{F}_{t^-}] = \frac{P_\theta[\mathcal{F}_{t^-}, v]}{\sum_{u \in \mathcal{U}^*(t^-)} P_\theta[\mathcal{F}_{t^-}, u]} = \frac{P_\theta[\mathcal{F}_{t^-}, v]}{L(\theta; t^-)} \tag{33}$$

and:

$$m^i_{v_i}(t^-) = \sum_{u_i = v_i} m_u(t^-) = \frac{\sum_{u_i = v_i} P_\theta[\mathcal{F}_{t^-}, v]}{L(\theta; t^-)}. \tag{34}$$

We must show that whatever is observed at time t — joining the pedigree, cases of onset, genetic tests, deaths or censored observations — and during periods between events, this happens in such a way that (30) always has the form of Elston's likelihood.

We suppose the pedigree genotype at time 0^- is $(0, 0, 0, \ldots)$, the only element of $\mathcal{U}^*(0^-)$, so:

$$L(\theta; 0^-) = \sum_{v \in \mathcal{U}^*(0^-)} m_v(0^-) = 1. \tag{35}$$

4.3. Contributions From New Members Joining

Next, suppose the ith member joins at time t (so $N(t) = i$) and assume, in inductive fashion, that $L(\theta; t^-)$ is as in (32), the sum being over $u' \in \mathcal{U}^*(t^-)$. We have:

$$L(\theta; t) = L(\theta; t^-)(d\mathbf{A}_0(t) P[e_0(t)])^{d\mathbf{N}_0(t)} = \sum_{u' \in \mathcal{U}^*(t^-)} P_\theta[\mathcal{F}_{t^-}, u'] \lambda_0(t) P[e_0(t)]. \tag{36}$$

The ith member may be a founder (a spouse marrying into the pedigree) assumed to join at some age $x_i(t) > 0$ or a non-founder (an offspring of pedigree members) who joins at age 0. If the ith member is a non-founder we assume that $P[e_0(t)]$ does not involve θ and:

$$L(\theta; t) \propto \sum_{u' \in \mathcal{U}^*(t^-)} P_\theta[\mathcal{F}_{t^-}, u'] \sum_{u_i = 1}^{u_i = M} P_\theta\left[u_i \mid u', e_0(t)\right] \tag{37}$$

$$= \sum_{u' \in \mathcal{U}^*(t^-)} \sum_{u_i = 1}^{u_i = M} P_\theta[\mathcal{F}_{t^-}, u'] \, p^i_u \tag{38}$$

$$= \sum_{u \in \mathcal{U}^*(t)} P_\theta[\mathcal{F}_t, u] \tag{39}$$

where: (a) the sum over u_i in (37) is identically 1; (b) we note that $P_\theta[u_i \mid u', e_0(t)] = p^i_u$; and (c) $u = (u'_1, \ldots, u'_{i-1}, u_i, 0, 0, \ldots)$. If the ith member

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

is a founder (recall that they enter state 1) then the mark $e_0(t)$ is their life history so far, so:

$$P[e_0(t)] \propto \sum_{u_i=1}^{u_i=M} \phi_{u_i} \exp\left(-\int_0^{x_i(t)} \mu_{u_i}(x)\,dx\right) \tag{40}$$

which is the probability that onset has not occurred by age $x_i(t)$, so p_u^i in (38) is replaced by:

$$\phi_{u_i} \exp\left(-\int_0^{x_i(t)} \mu_{u_i}(x))\,dx\right) = p_u^i \exp\left(-\int_0^{x_i(t)} \mu_{u_i}(x)\,dx\right). \tag{41}$$

The first member joining at time $t = 0$ is a special case. Also, the above is easily extended to multiple births, with $N(t) - N(t^-) > 1$.

4.4. Contributions From Periods With No Events

Suppose there is an event at time t^*, and then no event between time t^* and t. Noting that $\mathcal{R}_s = \mathcal{R}_t$ for $t^* < s \le t$, the relevant factors of the likelihood during $(t^*, t]$ can be written:

$$\prod_{i\in\mathcal{R}_t} \prod_{k\neq 3} \prod_{(t^*,t]} (1 - \lambda_{k3}^i(s)\,ds)^{\mathbf{Y}_{k3}^i(s)} = \prod_{i\in\mathcal{R}_t} \prod_{k\neq 3} \exp\left(-\int_{t^*}^{t} \lambda_{k3}^i(s)\,ds\right)^{\mathbf{Y}_{k3}^i(t)}. \tag{42}$$

For $t^* < s \le t$ the genotype probabilities evolve as follows (Bayes' Theorem):

$$m_v(s^-) = \frac{m_v(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{v_j}(x_j(r))\,dr\right)}{\sum_{u\in\mathcal{U}(t^*)} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)} \tag{43}$$

$$m_{v_i}^i(s^-) = \frac{\sum_{u_i=v_i} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)}{\sum_{u\in\mathcal{U}(t^*)} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)}. \tag{44}$$

Substitute this and (29) into (42), and the contribution is:

$$\prod_{i\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{t} \sum_{v_i=1}^{v_i=M} \frac{\sum_{u_i=v_i} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)}{\sum_{u\in\mathcal{U}(t^*)} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)} \mu_{v_i}(x_i(s))\,ds\right)$$

$$= \exp\left(-\int_{t^*}^{t} \sum_{i\in\mathcal{R}_t} \sum_{v_i=1}^{v_i=M} \frac{\sum_{u_i=v_i} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)}{\sum_{u\in\mathcal{U}(t^*)} m_u(t^*) \prod_{j\in\mathcal{R}_t} \exp\left(-\int_{t^*}^{s} \mu_{u_j}(x_j(r))\,dr\right)} \mu_{v_i}(x_i(s))\,ds\right)$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

$$= \exp\left(-\int_{t^*}^{t} \frac{\sum_{i\in\mathcal{R}_t}\sum_{v_i}\sum_{u_i=v_i} m_u(t^*)\prod_{j\in\mathcal{R}_t}\exp\left(-\int_{t^*}^{s}\mu_{u_j}(x_j(r))\,dr\right)\mu_{v_i}(x_i(s))}{\sum_{u\in\mathcal{U}(t^*)} m_u(t^*)\prod_{j\in\mathcal{R}_t}\exp\left(-\int_{t^*}^{s}\mu_{u_j}(x_j(r))\,dr\right)} ds \right)$$

$$= \exp\left(-\int_{t^*}^{t} \frac{\sum_{i\in\mathcal{R}_t}\sum_{u\in\mathcal{U}(t^*)} m_u(t^*)\prod_{j\in\mathcal{R}_t}\exp\left(-\int_{t^*}^{s}\mu_{u_j}(x_j(r))\,dr\right)\mu_{u_i}(x_i(s))}{\sum_{u\in\mathcal{U}(t^*)} m_u(t^*)\prod_{j\in\mathcal{R}_t}\exp\left(-\int_{t^*}^{s}\mu_{u_j}(x_j(r))\,dr\right)} ds \right)$$

$$= \exp\left(-\int_{t^*}^{t} -\frac{\partial}{\partial s}\log\left(\sum_{u\in\mathcal{U}(t^*)} m_u(t^*)\prod_{i\in\mathcal{R}_t}\exp\left(-\int_{t^*}^{s}\mu_{u_i}(x_i(r))\,dr\right)\right) ds \right)$$

$$= \sum_{u\in\mathcal{U}(t^*)} m_u(t^*)\prod_{i\in\mathcal{R}_t}\exp\left(-\int_{t^*}^{t}\mu_{u_i}(x_i(s))\,ds\right). \tag{45}$$

(Note that if the ith member has had a genetic test by time t^*, their genotype is known, fixed as g_i, say; all genotype probabilities $m_u(t^*)$ inconsistent with this vanish, so the numerators and denominators in the integrals above are equal; and the integrand reduces to $\mu_{g_i}(x_i(s))$.) Therefore, from (33), and making age instead of calendar time the variable of integration:

$$L(\theta;t) \propto \sum_{u\in\mathcal{U}^*(t^*)} P_\theta[\mathcal{F}_{t^*}, u]\prod_{i\in\mathcal{R}_t}\exp\left(-\int_{x_i(t^*)}^{x_i(t)}\mu_{u_i}(s)\,ds\right). \tag{46}$$

That is, between jumps, each member currently at risk of onset survives free of onset and contributes the probability of that event to the likelihood.

4.5. Contributions From Cases of Onset

If the ith member suffers onset at time t, $L(\theta;t)$ is:

$$\begin{aligned} L(\theta;t) &= L(\theta;t^-)\,d\mathbf{A}^i_{13}(t)^{d\mathbf{N}^i_{13}(t)} \propto L(\theta;t^-)\,\lambda^i_{13}(t) \\ &= L(\theta;t^-)\sum_{v_i=1}^{v_i=M} m^i_{v_i}(t^-)\mu_{v_i}(x_i(t)). \end{aligned} \tag{47}$$

Applying (34), and noting that $\mathcal{U}^*(t) = \mathcal{U}^*(t^-)$, this is:

$$\begin{aligned} \sum_{v_i=1}^{v_i=M}\sum_{u_i=v_i} P_\theta[\mathcal{F}_{t^-}, u]\,\mu_{v_i}(x_j(t)) &= \sum_{u\in\mathcal{U}^*(t^-)} P_\theta[\mathcal{F}_t, u]\,\mu_{u_i}(x_i(t)) \\ &= \sum_{u\in\mathcal{U}^*(t)} P_\theta[\mathcal{F}_t, u]. \end{aligned} \tag{48}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

(This is where the definition of $\mathcal{U}^*(t)$ in terms only of information gained from genetic tests matters.) Note that if either: (a) the ith member had previously had a genetic test; or (b) the disease has no cause except a single gene mutation; then the sum in (47) has only one non-zero term.

4.6. Contributions From Genetic Tests

If the ith member has a genetic test at time t, revealing their genotype to be v_i, then:

$$L(\theta;t) = L(\theta;t^-)\,(d\mathbf{A}^i_{12}(t)\,P[e^i_{12}(t) = v_i])^{d\mathbf{N}^i_{12}(t)} \propto L(\theta;t^-)\,\lambda^i_{12}(t)m^i_{v_i}(t^-). \quad (49)$$

Dropping $\lambda^i_{12}(t)$ (compare with Section 1.2, just after (7), where we assumed that the probability of being tested does not depend on θ) and applying (34):

$$L(\theta;t) \propto \sum_{\substack{u\in\mathcal{U}^*(t^-)\\ u_i = v_i}} P_\theta[\mathcal{F}_{t^-}, u] = \sum_{u\in\mathcal{U}^*(t)} P_\theta[\mathcal{F}_t, u]. \quad (50)$$

4.7. Contributions From Deaths and Censored Observations

Following death or loss to observation at time t, no further contributions are made to the likelihood.

4.8. Constructing the Likelihood

Let $x^*_i(T)$ represent the highest age at which the ith member has been observed alive and asymptomatic up to and including calendar time T. Starting with (35), applying (46) between events, and (38), (48) or (50) when an event other than censoring occurs, $L(\theta, T)$ is:

$$L(\theta;T) = \sum_{v\in\mathcal{U}^*(T)} \left(\prod_{i=1}^{i=N(T)} p^i_v \exp\left(-\int_0^{x^*_i(T)} \mu_{v_i}(s)\,ds\right) \prod_{\substack{\mathbf{N}^i_{13}(T)=1\\ \text{or } \mathbf{N}^i_{23}(T)=1}} \mu_{v_i}(x^*_i(T)) \right). \quad (51)$$

To within factors not depending on θ, this is the likelihood (12). It extends Elston's likelihood (9) allowing for genetic tests, and, despite appearances, we have not conditioned upon genotypes known at time T, as indeed we may not within the framework of survival analysis.

5. Retrospective Studies

5.1. Retrospective Studies, and Distorted Intensities and Probabilities

The model does not reflect the way in which real pedigrees are observed. Usually, 'interesting' families are identified (ascertained) through one or more affected

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

members, often called probands, and the pedigree is extended from them using medical records and/or examinations. In other words, pedigrees are not randomly sampled, they are selected because certain events have occurred. For the large literature on ascertainment bias see Cannings & Thompson (1977) and Hodge (2002). Elston (1973) did consider ascertainment bias, but not in a survival analysis framework.

There is not much literature on ascertainment bias in the context of survival models, see Li (2007). The construction of the likelihood (51) showed the essential rôle of time, and the need to work with a fully specified model. We now show that the fully specified model gives insight into the problem of ascertainment 'after the event'.

In Sections 3 and 4, it was clear what 'the' pedigree meant, namely that pedigree evolving forwards in time from a founding member. When ascertainment is retrospective, it is less clear what 'the' pedigree means, unless there are uniform rules about who is to be included in the pedigree, starting with those persons who directly caused 'it' to be ascertained. One simple rule, which we shall use for an example, is to limit each pedigree sampled to a single nuclear family, or to a sibship (all the children in a nuclear family). Cannings & Thompson (1977) gave more flexible rules in the form of sequential procedures.

Suppose the MPP representing a *given* pedigree is observed if and only if an event W occurs. Then the observed pedigree history is $\mathcal{G}_t = \mathcal{F}_t \vee W$, and the observed process is the MPP with 'distorted' intensities denoted as $\tilde{\lambda}_h(t,e)$ ($h \in \mathcal{H}^*$), (note that this is the only point at which we have needed to include the mark e in the intensity) and genotype probabilities denoted as $\tilde{m}_u(t) = P_\theta[u\,|\,\mathcal{G}_t]$, satisfying respectively:

$$\tilde{\lambda}_h(t,e) = \lambda_h(t,e)\,\frac{P[W\,|\,\mathcal{F}_{t^-}, d\mathbf{N}_h(t)=1, e_h(t)=e]}{P[W\,|\,\mathcal{F}_{t^-}]} \tag{52}$$

$$\tilde{m}_u(t) = m_u(t)\,\frac{P[W\,|\,u,\mathcal{F}_{t^-}]}{P[W\,|\,\mathcal{F}_{t^-}]}. \tag{53}$$

(This is along the same lines as Hoem (1969) and Aalen *et al.* (1980), who considered Markov processes, and also the distortion of counting process intensities in retrospective studies in Langholz *et al.* (1999).) Denote the fraction in (52) $D_h(t)$. If $D_h(t) > 1$ ($D_h(t) < 1$) then the event h makes W more (less) likely, hence the pedigree more (less) likely to be observed. Only if $D_h(t) = 1$ does the event h not matter. This could happen if W occurred before time t, in which case, omitting all observations before W occurred would give a partial likelihood.

5.2. Likelihoods

The likelihood is the probability of the process observed, as follows:

$$L(\theta;T) = \prod_{h\in\mathcal{H}^*}\prod_{[0,T]} (1-\mathbf{Y}_h(s)\,\tilde{\lambda}_h(s)\,ds)^{1-d\mathbf{N}_h(s)}(\mathbf{Y}_h(s)\,\tilde{\lambda}_h(s,e))^{d\mathbf{N}_h(s)}. \tag{54}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

In Section 4, we discarded nuisance parameters whenever an intensity or a mark probability not depending on θ appeared as a factor. Now, when we construct the likelihood sequentially, it is the distorted intensities and probabilities that appear as factors. These can be discarded only if the relevant $D_h(t)$ are rather simple: for example, if the ith member has a genetic test at time t, we can discard factor $\tilde{\lambda}^i_{12}(t)$ from the likelihood only if $D^i_{12}(t)$ does not involve the onset intensities $\lambda^j_{13}(t)$ and $\lambda^j_{23}(t)$ in respect of *any* pedigree member. This will not often be the case.

The very titles of papers by epidemiologists, such as Elston (1995) or Vieland & Hodge (1995), tell us that retrospective ascertainment of pedigrees is difficult. Properly designed studies are strongly advocated, but in their absence the data that are to hand will be analysed, often with some approximate adjustment for ascertainment bias. Actuaries have to use caution when applying these results to questions of insurance pricing; for example, Macdonald, Waters & Wekwete (2003) reduced published onset rates of breast and ovarian cancer associated with BRCA1 and BRCA2 mutations by 50% and 75%, to allow for possible unquantified ascertainment bias.

The discussion above shows that it would be very hard, at best, to estimate onset rates when retrospective ascertainment introduced distortions $D_h(t)$, different for every event in the pedigree history. From (52), the distortions affect anyone who *could* have led to the ascertainment, not just those who actually *did*. However, we can still gain insight from simple special cases.

5.3. Application to Sibships

Sibships are the classic units of analysis in genetic studies. Ascertainment through probands is allowed for by conditioning on the fact of ascertainment (Fisher, 1934; Morton, 1959). Adapting the notation from George & Elston (1991) slightly, a sibship is characterised by three quantities: the number of siblings c, the number of affected siblings a, and the number of probands $z (c \geq a \geq z)$. The simplest ascertainment model is that we include a sibship in the study if it has at least one proband, $z \geq 1$. If $P'[c]$ is the distribution of sibship sizes, then:

$$P[c,a,z \,|\, z \geq 1\,] = \frac{P'[c]\, P[a \,|\, c]\, P[z \,|\, c,a]}{\sum_{i=1}^{i=\infty} \sum_{j=0}^{j=1} P'[i]\, P[j \,|\, i]\, P[z \geq 1 \,|\, i,j]}. \tag{55}$$

If any sibling is affected with probability ζ, and any affected sibling becomes a proband with probability π (hence the name 'π-model'), and if $P[z \,|\, c,a] = P[z \,|\, a]$, this has the explicit form:

$$P[c,a,z \,|\, z \geq 1] = \frac{P'[c]\binom{c}{a}\binom{a}{z}\zeta^a(1-\zeta)^{c-a}\pi^z(1-\pi)^{a-z}}{1-\sum_{i=1}^{i=\infty} P'[i](1-\zeta\pi)^i}. \tag{56}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

Two special cases are often considered: (a) 'complete ascertainment' when $\pi = 1$, so all affected members become probands; and (b) 'single ascertainment' when $\pi \to 0$, so the probability of a sibship having more than one proband vanishes.

The corresponding scheme in our model is that the number of members of the pedigree at time T is $c + 2$ (c siblings and the parents). The parents are the 1st and 2nd members, and the siblings are the 3rd and subsequent members. Suppose the disease of interest is rare, dominant and has late onset, so that no sibling will suffer onset before the sibship is complete. Then, for sibling i, $\lambda^i_{13}(t) = 0$ before the sibship is complete, while afterwards c is known and part of $\mathcal{F}_{t^-}$, so in (52) we can condition on known sibship size. The simplest set-up is as follows.

(a) One parent is identified as a mutation carrier before the siblings are at risk.
(b) There is no genetic testing and no censoring (not even death).
(c) Any affected member becomes a proband with probability π, and ascertainment depends on there being at least one proband at time T.

Because of (a), the siblings' life histories are (conditionally) independent, and genotype probabilities are undistorted ($\tilde{m}_u(t) = m_u(t)$). Because of (b), genotype will not be revealed before onset. Because of (c), the event W is 'at least one proband, alive or dead, at time T'.

Let $\mathcal{R}_t = \{i : \mathbf{Y}^i_1(t) = 1\}$ be the set of siblings still at risk of onset at time t, and suppose $a(t)$ members are affected at time t. If ascertainment fails to take place through these affected members, which has probability, $(1-\pi)^{a(t)}$, it must occur through one or more of the siblings in $\mathcal{R}_t$ suffering onset in the future and then becoming a proband. So for $i \in \mathcal{R}_t$:

$$D^i_{13}(t) = \frac{(1-(1-\pi)^{a(t)+1}) + (1-\pi)^{a(t)+1}\left(\left\{1 - \prod_{j\in\mathcal{R}_t-\{i\}}\left[(1-\pi)+\pi\exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)\right]\right\}\right)}{(1-(1-\pi)^{a(t)}) + (1-\pi)^{a(t)}\left(1 - \left\{\prod_{j\in\mathcal{R}_t}\left[(1-\pi)+\pi\exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)\right]\right\}\right)}$$

$$= \frac{1-(1-\pi)^{a(t)+1}\left(\prod_{j\in\mathcal{R}_t-\{i\}}\left[(1-\pi)+\pi\exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)\right]\right)}{1-(1-\pi)^{a(t)}\left(\prod_{j\in\mathcal{R}_t}\left[(1-\pi)+\pi\exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)\right]\right)}. \tag{57}$$

The likelihood (54) with these distortion factors could be maximised numerically.

Under complete ascertainment ($\pi = 1$) $D^i_{13}(t) = 1$ if $a(t) > 0$, because ascertainment is guaranteed, and if $a(t) = 0$, the limit of (57) as $\pi \to 1$ gives:

$$D^i_{13}(t) = \frac{1}{1 - \prod_{j\in\mathcal{R}_t}\exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)} \tag{58}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

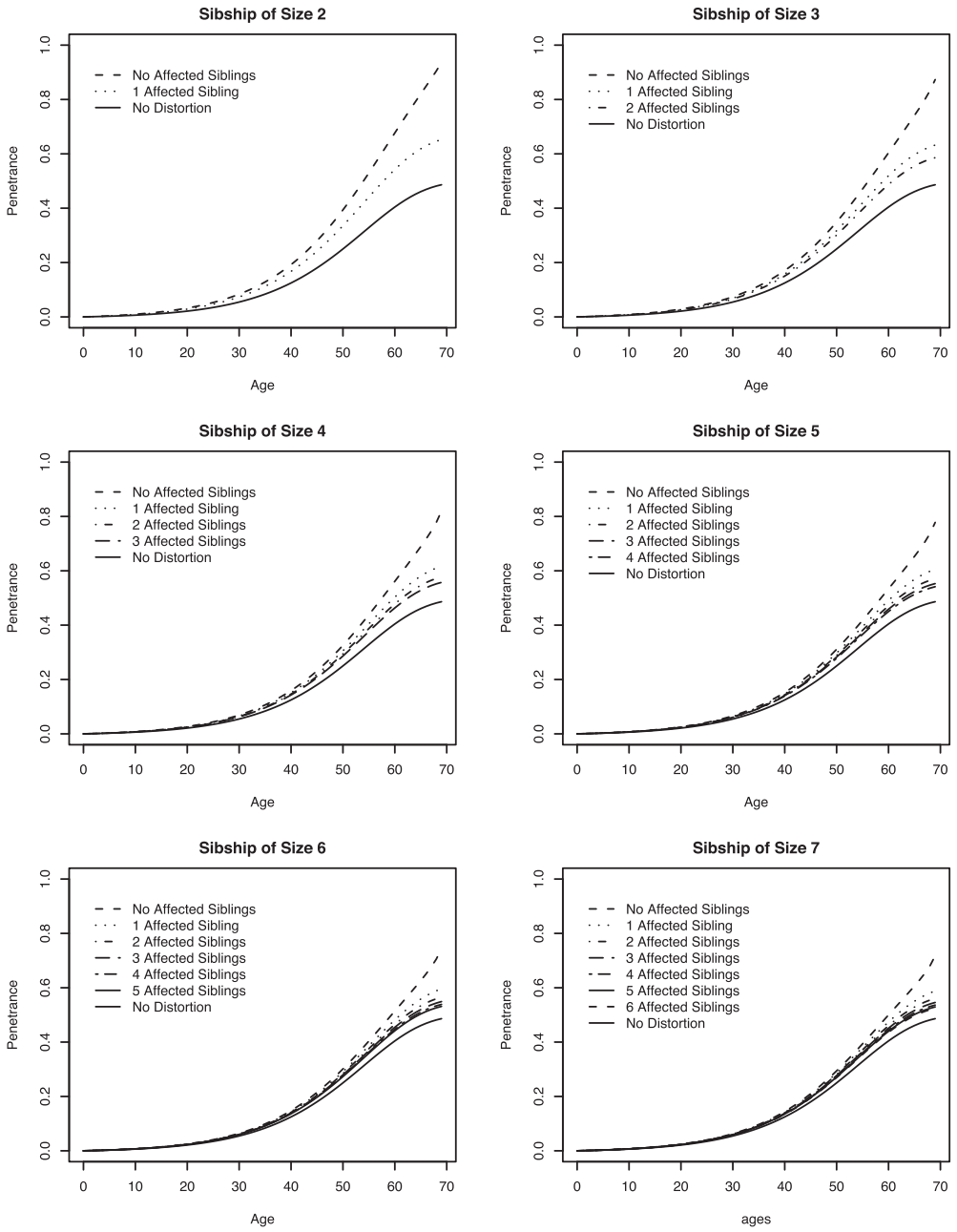


FIGURE 3: The penetrance as a function of age x under the observed process: $1 - \prod_{[0,x]}(1 - d\tilde{\lambda}^i_{13}(s))$, with single ascertainment. The actual rate of onset $\mu_2(x)$ is Gompertz: $\mu_2(x) = 0.000814 \exp(0.086178x)$. For sibships of size 2 to 7, the effect of the distorted intensities depends on the number of affected siblings. Also shown labelled 'No Distortion' is the penetrance based on the undistorted intensity $1 - \prod_{[0,x]}(1 - d\tilde{\lambda}^i_{13}(s))$.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

Under single ascertainment ($\pi \to 0$) both numerator and denominator of (57) become zero, but l'Hopital's rule gives:

$$D^i_{13}(t) = \frac{a(t) + 1 + \sum_{j \in \mathcal{R}_t - \{i\}} \left[1 - \exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)\right]}{a(t) + \sum_{j \in \mathcal{R}_t} \left[1 - \exp\left(-\int_t^T \lambda^j_{13}(s)\,ds\right)\right]}. \tag{59}$$

Under single ascertainment in the classical π-model, the probability that a sibship is ascertained is proportional to the number of affected siblings. The numerator and denominator of (59) are the expected numbers of affected siblings at time T, if $a(t) + 1$ or $a(t)$, respectively, are affected at time t, which we see is the corresponding property for age-at-onset data.

Figure 3 shows how the sibship size affects the distortion of the age-related penetrance under single ascertainment. The true rate of onset is Gompertz, with $\mu_2(x) = 0.000814 \exp(0.086178x)$. This results in actual penetrance of about 0.5 at age 50 among mutation carriers. Assuming the disorder to be rare and dominant, the penetrance in respect of siblings who are mutation carriers with probability 1/2 will approach 0.5 as age $x \to \infty$. The true penetrance is shown labelled 'No Distortion' and for comparison, the penetrance $1 - \prod_{[0,x]}(1 - d\tilde{\lambda}^i_{13}(s))$, assuming that $T = 70$ in (59), for sibships of size 2 to 7 (all siblings the same age, for simplicity) and for different numbers of affected siblings. The distorted penetrance is always overstated, and by more for smaller sibships.

If censoring (but not genetic testing) is introduced, with intensity $\tilde{\lambda}^i_{14}(t) = \lambda^i_{14}(t)\ D^i_{14}(t)$, (57) to (59) are changed only by replacing $(-\int_t^T \lambda^j_{13}(s)ds)$, the probability of not suffering onset, with the probability of not being *observed* to suffer onset:

$$\exp\left(-\int_t^T (\lambda^j_{13}(s) + \lambda^j_{14}(s))\,ds\right) + \int_t^T \exp\left(-\int_t^s (\lambda^j_{13}(r) + \lambda^j_{14}(r))\,dr\right)\lambda^j_{14}(s)\,ds \tag{60}$$

and, since if censoring occurs the person concerned will not be available as a proband, $D^i_{14}(t)$ is the same as $D^i_{13}(t)$ but with $a(t) + 1$ replaced by $a(t)$ whenever it appears in the numerator. Figure 4 shows the effect of adding censoring at rate $\lambda^i_{14}(t) = 0.01$ per year to the previous example. Comparing this with Figure 3 it is clear that the results are identical at $x = 70$ (as they must be) but that censoring increases slightly the observed penetrances at younger ages.

If we introduce genetic testing in the absence of censoring, we must distinguish between persons in States 1 and 2, : let $\mathcal{R}^1_t = \{i : \mathbf{Y}^i_{13}(t) = 1\}$ and $\mathcal{R}^2_t = \{i : \mathbf{Y}^i_{23}(t) = 1\}$. Then for $i \in \mathcal{R}^1_t$, $D^i_{13}(t)$ is:

$$\frac{1 - (1-\pi)^{a(t)+1}\left(\prod_{j \in \mathcal{R}^1_t - \{i\}} \left[(1-\pi) + \pi G^j_1(t)\right] \prod_{j \in \mathcal{R}^2_t} \left[(1-\pi) + \pi G^j_2(t)\right]\right)}{1 - (1-\pi)^{a(t)}\left(\prod_{j \in \mathcal{R}^1_t} \left[(1-\pi) + \pi G^j_1(t)\right] \prod_{j \in \mathcal{R}^2_t} \left[(1-\pi) + \pi G^j_2(t)\right]\right)} \tag{61}$$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

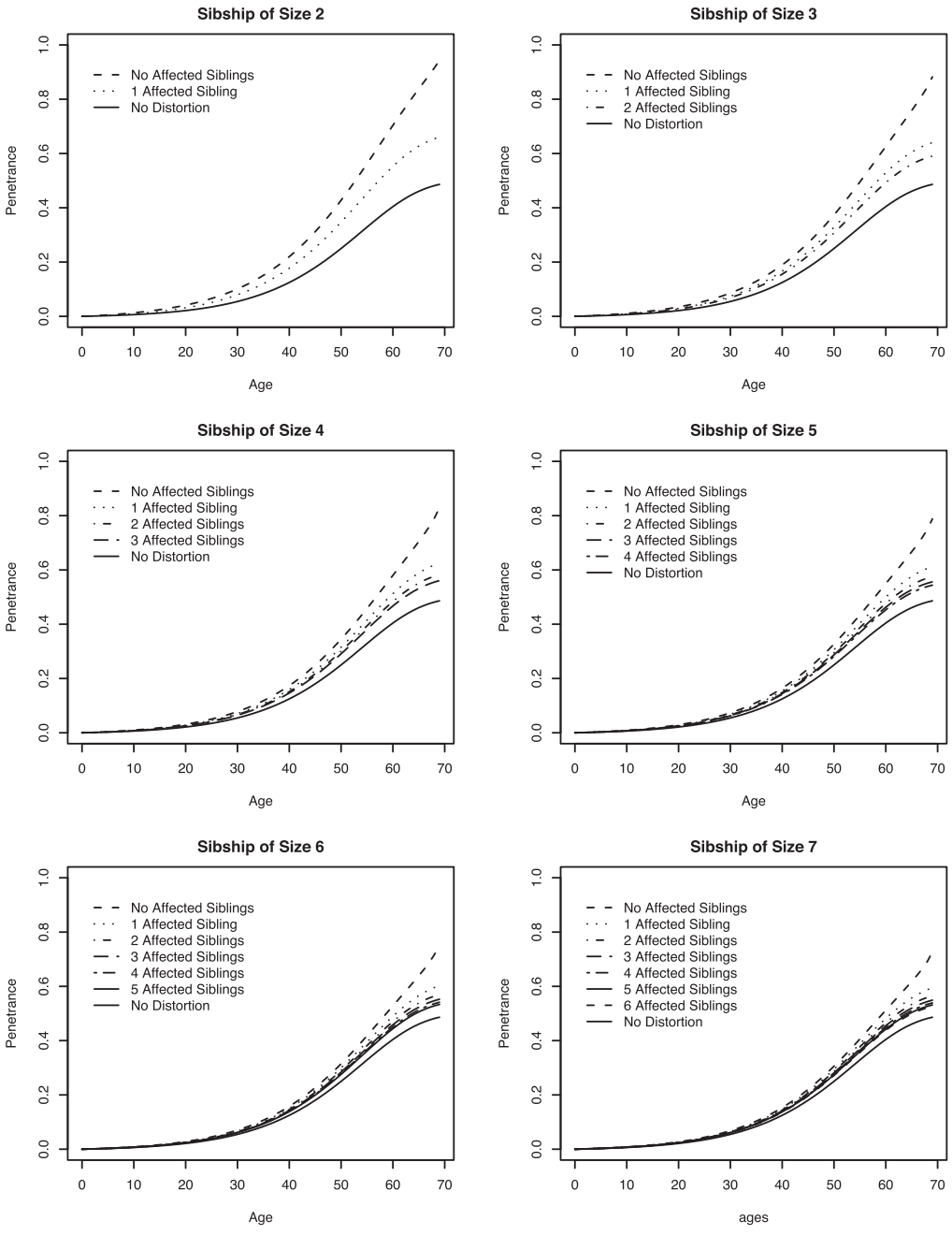


FIGURE 4: The penetrance as a function of age x in the presence of censoring, under the observed process: $1 - \prod_{[0,x]}(1 - d\tilde{\lambda}^i_{13}(s))$, with single ascertainment. The actual rate of censoring is $\lambda^j_{14}(t) = 0.01$ per year, and all other details are as in Figure 3.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

where:

$$G_1^j(t) = \exp\Big(-\int_t^T (\lambda_{12}^j(s) + \lambda_{13}^j(s))\,ds\Big) + \int_t^T \exp\Big(-\int_t^s (\lambda_{12}^j(r) + \lambda_{13}^j(r))\,dr\Big)\lambda_{12}^j(s)\exp\Big(-\int_s^T \lambda_{23}^j(r)\,dr\Big)ds \quad (62)$$

$$G_2^j(t) = \exp\Big(-\int_t^T \lambda_{23}^j(s)\,ds\Big) \quad (63)$$

and $D_{12}^i(t)$ is:

$$\frac{1-(1-\pi)^{a(t)}\Big(\prod_{j\in\mathcal{R}_t^1-\{i\}}\big[(1-\pi)+\pi G_1^j(t)\big]\prod_{j\in\mathcal{R}_t^2\cup\{i\}}\big[(1-\pi)+\pi G_2^j(t)\big]\Big)}{1-(1-\pi)^{a(t)}\Big(\prod_{j\in\mathcal{R}_t^1}\big[(1-\pi)+\pi G_1^j(t)\big]\prod_{j\in\mathcal{R}_t^2}\big[(1-\pi)+\pi G_2^j(t)\big]\Big)}. \quad (64)$$

Note that if the mutation is the sole cause of the disease, then for non-mutation carriers, $G_2^j(t) = 1$. For $i \in \mathcal{R}_t^2$, $D_{23}^i(t)$ is:

$$\frac{1-(1-\pi)^{a(t)+1}\Big(\prod_{j\in\mathcal{R}_t^1}\big[(1-\pi)+\pi G_1^j(t)\big]\prod_{j\in\mathcal{R}_t^2-\{i\}}\big[(1-\pi)+\pi G_2^j(t)\big]\Big)}{1-(1-\pi)^{a(t)}\Big(\prod_{j\in\mathcal{R}_t^1}\big[(1-\pi)+\pi G_1^j(t)\big]\prod_{j\in\mathcal{R}_t^2}\big[(1-\pi)+\pi G_2^j(t)\big]\Big)}. \quad (65)$$

Under single ascertainment, we obtain expressions for each of the above exactly like (59), of the form:

$$\frac{\text{Expected number of cases at time } T \text{ given } \mathcal{F}_{t^-} \text{ and if transition occurs}}{\text{Expected number of cases at time } T \text{ given } \mathcal{F}_{t^-}}. \quad (66)$$

It is not difficult also to write down expressions for all the $D_{kl}^i(t)$ when we restore both censoring and genetic testing to the model, but we omit them for brevity. Our main point is already clear from the forms of the $D_{kl}^i(t)$ above, namely that censoring and genetic testing cannot now be factored out of the likelihood. This is plain, since intensities other than $\tilde{\lambda}_{13}^i(t)$ and $\tilde{\lambda}_{23}^i(t)$ now depend on θ. It is not possible, therefore, to work with the partially specified model. Espinosa & Macdonald (2007) reached an analagous conclusion in a non-parametric setting, where even genuinely independent censoring, if it took place before sibships were ascertained, introduced bias.

A popular 'adjustment' for ascertainment is to omit the proband(s) from the likelihood (see, for example, Newcombe (1981)). Intuitively, excluding those persons whose lifetimes, observed retrospectively, led to the ascertainment,

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

ought to remove the 'contamination'. However, this is not correct, as it ignores the probability of other events that could have led to ascertainment. In a few special cases, 'omitting the proband' is correct, but as an algebraic quirk in a fully specified model (Cannings & Thompson, 1977; Thompson, 1993). That is not the case here. There seems to be no obvious way in practice to adjust the likelihood (54) to obtain, even approximately, the same estimates as the 'correct' likelihood (51).

6. Conclusions

Actuarial interest in evidence-based approaches to medical underwriting is growing, notably in respect of genetic information. This led us to look more closely at Elston's likelihood (9) usually used for age-at-onset estimation with pedigree data. We asked three questions, reproduced here:

(a) Can we recover Elston's likelihood from a properly formulated survival model?
(b) How are internal covariates generated by presymptomatic genetic tests to be included in the model?
(c) How does retrospective ascertainment: (i) interact with censoring; and (ii) affect estimates of onset rates?

We have answered the first two of these, by specifying a full MPP model of an entire pedigree, in the spirit of Arjas & Haara (1984). By constructing the likelihood sequentially in calendar time, we recovered Elston's likelihood, genetic tests included, showing: (a) that it is correct to use the partial model upon which it is based; and (b) the apparent conditioning on known genotypes revealed by genetic tests — problematic because they are internal covariates — in fact does not occur provided the full model is used. If sampling is retrospective and non-random, the observed process, hence the likelihood, is generated by distorted versions of the 'true' transition intensities. We observed the following:

(a) In general, the distorted intensities are functions of time (hence age) and are different for each member of the pedigree.
(b) The distortion factors are functions of the transition intensities we are trying to estimate.
(c) We gave some examples of distorted onset intensities in the simplest setting (sibships), showing that nuisance parameters would not factorise out of the likelihood as before. Therefore, the partial model cannot be used correctly in this setting.
(d) We found no obvious way to adjust the likelihood to account for this, so intimately conjoined were the true rates of onset and the factors distorting them.
(e) We found in (59) and, more generally, (66), the age-at-onset counterparts for the well-known result that, under single ascertainment of sibships, the

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

probability of ascertainment is proportional to the number of affected siblings. This appears to be a new result.

(f) In particular, the popular adjustment of 'omit the proband(s)' had no justification here.

This tends to support the views of those genetic epidemiologists who have described the ascertainment problem as intractable; see Vieland & Hodge (1995). Our model shows that the special rôle of time in survival analysis introduces new problems in respect of pedigree data.

What should be done by the actuary who wishes to apply epidemiological studies to pricing, reserving and adverse selection problems? Or the regulator, who needs to test the evidence base for differential pricing? Retrospective studies are likely to yield onset rates with an unknown degree of ascertainment bias. This will be true of non-genetic studies as well as pedigree analyses. (This is, of course, the very reason why retrospective data are so often analysed by case-control studies; the estimated odds ratio *is* unbiased. Unfortunately, onset rates are of much more use to actuaries.) We can only conclude that actuaries must highlight the possible lack of robustness in such models, and even quite crude sensitivity tests (such as the arbitrary reductions in onset rates in Macdonald, Waters & Wekwete (2003)) might be advisable. It is then up to assessors, such as GAIC in the UK, to judge the evidence.

Acknowledgements

This work was carried out at the Genetics and Insurance Research Centre at Heriot-Watt University. We would like to thank the sponsors for funding, and members of the Steering Committee for helpful comments at various stages. We thank an anonymous referee for very constructive comments.

References

Aalen, O.O., Borgan, Ø., Keiding, N. and Thormann, J. (1980) Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring, *Scandinavian Journal of Statistics*, **7**, 161-171.

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993) *Statistical models based on counting processes*, Springer-Verlag, New York.

Arjas, E. (1989) Survival models and martingale dynamics, *Scandinavian Journal of Statistics*, **16**, 177-225.

Arjas, E. and Haara, P. (1984) A marked point process approach to censored failure data with complicated covariates, *Scandinavian Journal of Statistics*, **11**, 193-209.

Cannings, C. and Thompson, E.A. (1977) Ascertainment in the sequential sampling of pedigrees, *Clinical Genetics*, **12**, 208-212.

Elston, R.C. (1973) Ascertainment and age at onset in pedigree analysis, *Human Heredity*, **23**, 105-112.

Elston, R.C. (1995) 'Twixt cup and lip: How intractable is the ascertainment problem?, *American Journal of Human Genetics*, **56**, 15-17.

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

ESPINOSA, C. and MACDONALD, A.S. (2007) A correction for ascertainment bias in estimating rates of onset of highly penetrant genetic disorders, *ASTIN Bulletin*, **37**, 429-452.
FISHER, R.A. (1934) The effects of methods of ascertainment upon the estimation of frequencies, *Annals of Human Genetics*, **50**, 399-402.
GEORGE, V.T. and ELSTON, R.C. (1991) Ascertainment: An overview of the classical segregation analysis model for independent sibships, *Biometric Journal*, **33**, 741-753.
HODGE, S.E. (2002) Ascertainment, *Biostatistical genetics and genetic epidemiology*, Elston, R., Olson, J. and Palmer, L., John Wiley.
HOEM, J.M. (1969) Purged and partial Markov chains, *Skandinavisk Aktuarietidskrift*, **1969**, 146-155.
KALBFLEISCH, J.D. and PRENTICE, R.L. (2002) *The statistical analysis of failure time data (second edition)*, John Wiley, New Jersey.
LANGHOLZ, B., ZIOGAS, A., THOMAS, D.C., FAUCETT, C., HUBERMAN, M. and GOLDSTEIN, L. (1999) Ascertainment correction in rate ratio estimation from case-sibling control studies of variable age-at-onset diseases, *Biometrics*, **55**, 1129-1136.
LI, H. (2007) Survival analysis methods in genetic epidemiology, *Current topics in human genetics: Studies of complex diseases*, Deng, H.-W., Shen, H., Liu Y. and Hu, H., World Scientific Publishing, Singapore.
LU, L., MACDONALD, A.S. and WATERS, H.R. (2008) Sampling distributions of critical illness insurance premium rates: Breast and ovarian cancer, *ASTIN Bulletin*, **38**, 527-542.
LU, L., MACDONALD, A.S. and WEKWETE, C.T. (2008) Premium rates based on genetic studies: How reliable are they?, *Insurance: Mathematics and Economics*, **42**, 319-331.
MACDONALD, A.S., WATERS, H.R., and WEKWETE, C.T. (2003) The genetics of breast and ovarian cancer II: A model of critical illness insurance, *Scandinavian Actuarial Journal*, **1**, 28-50.
MORTON, N.E. (1959) Genetic tests under incomplete ascertainment, *American Journal of Human Genetics*, **11**, 1-16.
NEWCOMBE, R.G. (1981) A life table for onset of Huntington's Chorea, *Annals of Human Genetics*, **45**, 375-385.
THOMPSON, E.A. (1993) Sampling and ascertainment in genetic epidemiology: A tutorial review, *Technical Report 243, Department of Statistics, University of Washington.*
VIELAND, V.J. and HODGE, S.E. (1995) Inherent intractability of the ascertainment problem for pedigree data: A general likelihood framework, *American Journal of Human Genetics*, **56**, 33-43.

ANGUS MACDONALD
Department of Actuarial Mathematics and Statistics,
and the Maxwell Institute for Mathematical Sciences,
Heriot-Watt University,
Edinburgh EH14 4AS,
United Kingdom
Tel.: +44(0)131-451-3209
Fax: +44(0)131-451-3249
E-mail: A.S.Macdonald@ma.hw.ac.uk

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

NOTATION

b_i	calendar time of birth of ith pedigree member
$e_h(t)$	mark associated with event h at time t: $e_h(t) \in E_h(t)$
g	an individual person's genotype
h	event in $\mathcal{H}$ or $\mathcal{H}^*$: $h = (i,k,l)$ means ith person jumps from State k to State l
i, j	persons or pedigree members
k, l	states: events in a life history are represented by transitions between states
$m_u(t)$	probability of pedigree genotype u given observed information at time t
$m_g^i(t)$	probability that ith member has genotype g given observed information at time t
p	genotype transmission probabilities
r, s, t	calendar time
u, v	the joint genotype of the members of a pedigree: elements of $\mathcal{U}(t)$ or $\mathcal{U}^*(t)$
u_i, v_i	the genotype of the ith member of a pedigree: a component of u or v
x, y	age
$x_i(t)$	age of the ith pedigree member at calendar time t
(i,k,l)	the event that the ith pedigree member jumps from State k to State l
$D_h(t)$	distortion applied to intensity $\lambda_h(t)$
$D_{kl}^i(t)$	equal to $D_h(t)$ for event $h = (i,k,l)$; distortion applied to $\lambda_{kl}^i(t)$
$E_h(t)$	mark space associated with event h at time t: $e_h(t) \in E_h(t)$
$L(\theta, t)$	Likelihood for parameter θ based on observations up to and including time t
M	number of distinct genotypes
$N(t)$	number of pedigree members who have ever lived, as at calendar time t
T	calendar time when observation ceases; observation is in the interval $[0, T]$ of calendar time
$\mathbf{A}_h(t)$	compensator of counting process $\mathbf{N}_h(t)$
$\mathbf{A}_{kl}^i(t)$	equal to $\mathbf{A}_h(t)$ for event $h = (i,k,l)$
$\mathbf{N}_h(t)$	process counting occurrences of event h by time t
$\mathbf{N}_{kl}^i(t)$	equal to $\mathbf{N}_h(t)$ for event $h = (i,k,l)$
$\mathbf{Y}_h(t)$	process indicating being at risk of event h at time t^-
$\mathbf{Y}_{kl}^i(t)$	equal to $\mathbf{Y}_h(t)$ for event $h = (i,k,l)$

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press

$\mathcal{A}_t$	information known at time t given only the pedigree structure
$\mathcal{F}_t$	information known at time t given entire history
$\mathcal{H}$	space of possible events befalling pedigree members
$\mathcal{H}^*$	space of possible events befalling pedigree members, augmented by joining the pedigree
$\mathcal{R}_t$	risk set: set of persons at risk of a specified event or set of events
$\mathcal{U}(t)$	set of possible pedigree genotypes given pedigree at time t
$\mathcal{U}^*(t)$	$\mathcal{U}(t)$ excluding pedigree genotypes ruled out by genetic tests
ϕ_g	population frequency of genotype g
$\lambda_h(t)$	intensity of event h at time t, a function of calendar time
$\lambda^i_{kl}(t)$	equal to $\lambda_h(t)$ for event $h = (i,k,l)$; intensity of person-specific events
$\tilde{\lambda}_h(t)$	distorted intensity of event h at time t
$\tilde{\lambda}^i_{kl}(t)$	equal to $\tilde{\lambda}_h(t)$ for event $h = (i,k,l)$
$\mu_g(x)$	intensity of onset, a function of genotype g and age x: the target of estimation
θ	parameter of the genetic model
π	probability that an affected sibling becomes a proband, in the π-model
ζ	probability that a sibling is affected, in the π-model

https://doi.org/10.2143/AST.40.1.2049218 Published online by Cambridge University Press