Survival Analysis on Pedigrees: A Marked Point Process Model

Angus Macdonald

Heriot-Watt University
and the Maxwell Institute for Mathematical Sciences
Outline

1. Introduction
2. Genetic tests introduce internal covariates
3. Pedigree likelihoods with genetic testing
4. Retrospective ascertainment
5. An example using sibships
Introduction

Problem: How to assess the cost of insurance in the presence of genetic information?

The purpose may be:

- To price an insurance contract.
- To assess the adverse selection risk if genetic information can’t be used.

Genetic information may mean:

- A family history of an inherited disease.
- The result of a DNA-based genetic test.
Critical Illness (Dread Disease) Insurance

Figure 1: A multiple state model for genetic disease $X$ in Critical Illness insurance.
Figure 2: A hypothetical example of a pedigree. Squares are males, circles are females, and a slash denotes death. Affected individuals are shown as filled squares/circles.
Survival Analysis on Pedigrees

We consider a rare Mendelian dominant disorder \( \Rightarrow \) an affected parent passes on the ‘risky’ gene to each child with probability 1/2.

Also, the disease has no causes except the ‘risky’ gene (e.g. Huntington’s disease, polycystic kidney disease).

Three features that characterize survival analysis on pedigrees:

1. Pedigree members have dependent lifetimes because they share genes. *What is the form of the likelihood?*

2. Pedigrees are often selected *because* they contain affected members. *How can we allow for non-random ascertainment?*

3. Censoring affects the probability of ascertainment. *How do censoring and ascertainment interact?*
Genetic Tests Yield Internal Covariates

Consider a person alive and unaffected by a disorder at age 40. What do they contribute to the likelihood of onset/survival at age 30 (say)?

**Question 1:** If their genotype is unknown?

**Answer:** \( P[\text{Mutation}] \times P[\text{No Onset at Age 30} | \text{Mutation}] \)

**Question 2:** Known to be a mutation carrier?

**Answer:** \( P[\text{No Onset at Age 30} | \text{Mutation}] \)

**Question 3:** Known to be a mutation carrier because they took a presymptomatic genetic test at age 40?

**Answer:** \( P[\text{No Onset at Age 30} | \text{Presymptomatic Test at Age 40}] = 1 \)

The genotype is an internal covariate.
A Closer Look at Internal Covariates

- ‘Events’ in a life history are $A_1, A_2, \ldots, A_n$ in time order.
- Observed covariate is $v \in \mathcal{U}$.
- Model parameter is $\theta$.

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

But, how did we learn about $v$?
A Closer Look at Internal Covariates

We must have observed an event $C$ that revealed it. Suppose we observe $C$ between $A_j$ and $A_{j+1}$. The likelihood is:

$$L(\theta; n) = P[A_1] P[A_2|A_1] \cdots P[A_j|A_1, \ldots, A_{j-1}] \times P[C|A_1, \ldots, A_j]$$

$$\times P[A_{j+1}|C] P[A_{j+2}|C, A_{j+1}] \cdots P[A_n|C, A_{j+1}, \ldots, A_{n-1}]$$

So, condition on $v$ only after $C$ is observed?

Decompose $C$ into an event $C_j$ — the test occurring at a random time $j$ say — and the information or mark — the genotype $v$ — obtained at time $j$ (Arjas & Haara, 1984).
A Closer Look at Internal Covariates

Likelihoods just before and after the test $C_j$ at time $j$ are:

$$L(\theta; j^-) = \sum_{u \in \mathcal{U}} P[u] P[A_1|u] \cdots P[A_j|u, A_1, \ldots, A_{j-1}]$$

$$L(\theta; j^+) = L(\theta; j^-) P[C_j|A_1, \ldots, A_j] P[v|C_j, A_1, \ldots, A_j].$$

Assume:

- $P[C_j|A_1, \ldots, A_j]$ does not depend on $\theta$.
- $P[v|C_j, A_1, \ldots, A_j] = P[v|A_1, \ldots, A_j]$. 
A Closer Look at Internal Covariates

Then:

\[ L(\theta; j^+) \propto L(\theta; j^-) P[v|A_1, \ldots, A_j] \]
\[ = L(\theta; j^-) \frac{P[v] P[A_1|v] P[A_j|v, A_1, \ldots, A_{j-1}]}{\sum_{u \in \mathcal{U}} P[u] P[A_1|u] \cdots P[A_j|u, A_1, \ldots, A_{j-1}]} \]
\[ = P[v] P[A_1|v] P[A_j|v, A_1, \ldots, A_{j-1}]. \]

Arjas & Harra (1984): only by looking at the full model, including \( C_j \) as well as \( v \), can we decide whether we can use the partial model for inference on \( \theta \).
Pedigrees and Elston’s Likelihood

Figure 3: A hypothetical example of a pedigree. Squares are males, circles are females, and a slash denotes death. Affected individuals are shown as filled squares/circles.
Pedigrees and Elston’s Likelihood

A genetic model has parameter $\theta$ which includes:

- The set of distinct individual genotypes, labelled $\{1, 2, \ldots, M\}$.
- The genotype frequencies $\phi_g$ in the population.
- Genotype-specific hazard rates:

  $$\mu_g(x) = \text{Rate of disease onset at age } x.$$  

Note that we may have $\int_0^{\infty} \mu_g(x) dx < \infty$, called ‘incomplete penetrance’

A given pedigree with $N$ members has joint genotype $u = (u_1, u_2, \ldots, u_N) \in \mathcal{U} = \{1, 2, \ldots, M\}^N$. Given $\theta$ we can find $P_\theta[u]$.
Pedigrees and Elston’s Likelihood

Elston (1973) obtained the pedigree likelihood:

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

where the \( j \)th person is observed up to age \( x_j^* \) and \( \mathbb{I}_j \) is an indicator that observation ended with disease onset.

It is defined in terms of all that is known retrospectively at the time of analysis and in terms of age-specific hazards of disease onset.

Presymptomatic genetic testing did not exist in 1973.
Elston’s Likelihood and Genetic Testing

Define:

\[ \mathcal{U} = \text{Set of all possible pedigree genotypes} \]
\[ \mathcal{U}^* = \text{Set of all pedigree genotypes consistent with known presymptomatic test results} \]

Then we would hope that the likelihood allowing for genetic testing would be:

\[
L_1(\theta) = \sum_{u \in \mathcal{U}^*} \prod_{j=1}^{j=N} \left[ \exp \left( - \int_0^{x_j^*} \mu_{u,j}(x, \theta) \, dx \right) \mu_{u,j}(x_j^*, \theta)^{I_j} \right] P_\theta[u]
\]

But presymptomatic testing yields internal covariates
Figure 4: Model of life history of the $i$th 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) $\lambda_{jk}^i(t)$ depending on calendar time $t$. 
A Marked Point Process Model

Usual counting process definitions for a transition between states $j$ and $k$ by the $i$th person at calendar time $t$.

$$N_{jk}^i(t) = \text{No. transitions } j \rightarrow k \text{ by } i\text{th person}$$
$$e_{jk}^i(t) = \text{Mark obtained on transition}$$
$$Y_{jk}^i(t) = \text{Indicates } i\text{th person at risk at time } t^-$$
$$\lambda_{jk}^i(t) = \text{Hazard rate at time } t$$
$$A_{jk}^i(t) = \int_0^t Y_{jk}^i(s)\lambda_{jk}^i(s)ds$$

Model defined in calendar time. If $i$th person is born at time $b_i$ their age at time $t$ is $x_i(t) = t - b_i$. Must model all pedigree members simultaneously since their life histories are dependent.
A Marked Point Process Model

Likelihood based on observation up to time $T$ is:

$$L(\theta, T) = \prod_{h \in H} \prod_{[0, T]} \left(1 - dA_h(t)\right)^{1-dN_h(t)} \left(dA_h(t) P[e_h(t)|F_{t-}, dN_h(t) = 1]\right)^{dN_h(t)}.$$ 

where $H$ is the set of all triples $(i, j, k)$.

Note events at time $t$ are conditioned only on $F_{t-}$, the history before calendar time $t$. So for histories like:

- $X$ dies in 1950.
- $X$’s grandson has a genetic test in 2010, fixing genotype as $g$

the genotype cannot enter the likelihood until 2010, although it carries information about $X$. How does this happen?
The Problem

Two likelihoods:

\[
\sum_{u \in U^*} \prod_{j=1}^{j=N} \left[ \exp \left( - \int_{0}^{x_j^*} \mu_{u_j}(x, \theta) \, dx \right) \mu_{u_j}(x_j^*, \theta)^{I_j} \right] P_{\theta}[u]
\]

conditioning on everything known at time \( T \) and based on age-related hazard rates; and

\[
L(\theta, T) = \prod_{h \in H} \prod_{[0,T]} (1 - dA_h(t))^{1-dN_h(t)} (dA_h(t) P[e_h(t)|F_{t-}, dN_h(t) = 1])^{dN_h(t)}
\]

Conditioning events at time \( t \) only on \( F_{t-} \) and based on transition intensities in calendar time.

Are they equivalent?
Outline of an Answer

1. Define joint counting process model of entire pedigree.
   - Start with one founder member.
   - Add a new event to the model — new member joins the pedigree. At time $t$ there are $N(t)$ members.
   - Pedigree genotypes at time $t$ have the form:

   \[
   u = \{u_1, \ldots, u_{N(t)}, 0, 0, \ldots\} \in \mathcal{U}^*(t).
   \]

   - New member joining at $t$ expands $\mathcal{U}^*(t^-)$ to a larger $\mathcal{U}^*(t)$.
   - Genetic test at $t$ shrinks $\mathcal{U}^*(t^-)$ to a smaller $\mathcal{U}^*(t)$. 
Outline of an Answer

2. Define marks.

- When a genetic test occurs, the mark is the **genotype** of the tested person.

- When a new member joins the pedigree, the mark is:
  - **age, sex and spouse**, if they marry into it (‘founders’); or
  - **who the parents are**, if they are born into it (‘joiners’).

- The mark probabilities enter the product integral likelihood at time $t$ but not before time $t$.

- This is the essential difference between a model of the pedigree and separate models of its members.
Outline of an Answer

3. Define information and genotype probabilities.

- Filtration based on all information at time \( t \) is \( \mathcal{F}_t \).
- This filtration appears in the product integral likelihood.
- Define genotype probability \( m_u(t) = P_{\theta}[u|\mathcal{F}_t] \).
- Define \textit{individual} genotype probability \( m^i_g(t) = \sum_{u_i=g} m_u(t) \).
- Filtration based solely on the family structure at time \( t \) is \( \mathcal{A}_t \).
- Probabilities \( P_{\theta}[u] \) in Elston's likelihood use \( \mathcal{A}_T \).
- Define genotype probability \( p_u(t) = P_{\theta}[u|\mathcal{A}_t] \).
Outline of an Answer

4. Construct the product integral likelihood sequentially.

- Note that genotype probabilities are related to the likelihood via:

\[ m_v(t^-) = P_{\theta}[v|F_{t^-}] = \frac{P_{\theta}[F_{t^-}, v]}{\sum_{u \in U^*(t^-)} P_{\theta}[F_{t^-}, u]} = \frac{P_{\theta}[F_{t^-}, v]}{L(\theta; t^-)} \]

- Note that, in between transitions, genotype probabilities evolve:

\[ m_v(s^-) = \frac{m_v(t^*) \prod_{j \in R_t} \exp \left( - \int_{t^*}^s \mu_{v_j}(x_{j}(r)) \, dr \right)}{\sum_{u \in U(t^*)} m_u(t^*) \prod_{j \in R_t} \exp \left( - \int_{t^*}^s \mu_{u_j}(x_{j}(r)) \, dr \right)} \]

where the last transition was at time \( t^* \) and the set of persons at risk after the last transition is \( R_t \).

- Connect these with the mark probabilities at event times, and we can show the two likelihoods are equivalent (see Macdonald, 2010).
Outline of an Answer

4. Construct the product integral likelihood sequentially.

Product integral likelihood is (proportional to):

$$\sum_{v \in U^*(T)} \left( \prod_{i=1}^{i=N(T)} p^i_v \exp \left( - \int_0^{x_i^*(T)} \mu_{v_i}(s) ds \right) \prod_{N_{13}^i(T) = 1 \text{ or } N_{23}^i(T) = 1} \mu_{v_i}(x_i^*(T)) \right).$$

- We have introduced genetic testing to Elston’s likelihood.
- We need to assume that the probability of being tested does not depend on $\theta$.
- There is no violation of information.
- But we have not allowed for retrospective ascertainment.
Ascertainment

Pedigrees are not observed forwards in time starting with a first founder.

Pedigrees come to the attention of researchers because they contain affected members. This is called ascertainment.

This is a source of bias that has long been a problem in pedigree studies.


But not much has been written about survival analysis for time-to-onset data on pedigrees.


**Ascertainment**

Suppose a *given* pedigree is ascertained if and only if an event $W$ occurs.

This is a simplification — it assumes that pedigrees are waiting to be ascertained.

The effect is to **distort** the observed intensities and genotype probabilities. The observed filtration is now $G_t = \mathcal{F}_t \vee W$ and the observed quantities are:

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

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

The Distorted Likelihood

The likelihood using the distorted intensities is:

\[
\prod_{h \in H^*} \prod_{[0,T]} \left( 1 - Y_h(s) \tilde{\lambda}_h(s) \, ds \right)^{1-dN_h(s)} \left( Y_h(s) \tilde{\lambda}_h(s, e) \right)^{dN_h(s)}.
\]

Define:

\[
D_h(t, e) = \frac{P[W | \mathcal{F}_{t-}, dN_h(t) = 1, e_h(t) = e]}{P[W | \mathcal{F}_{t-}]}.
\]

- Estimated onset rates \( \tilde{\lambda}_{jk}^i(t) \) will be biased unless \( D_{jk}^i(t) = 1 \).

- Nuisance onset rates, e.g. \( \lambda_{12}^i(t) \), cannot now be dropped from the likelihood.
An Example Based on Probands

The classical pedigree is the nuclear family, parents and children. It may be ascertained via the children as follows.

To be included in a study:

- there must be $c > 0$ siblings;
- of whom $a > 0$ are affected; and
- at least $z > 0$ affected cases cause the pedigree to be included.

The $z$ cases who bring the study to the researchers are called probands. The probability that a case becomes a proband is $\pi$ (hence $\pi$-model). Extreme cases are:

- $\pi = 1$, hence all cases become probands — complete ascertainment.
- $\pi \to 0$, hence only one proband per pedigree — single ascertainment.
Figure 5: A hypothetical example of a pedigree. Squares are males, circles are females, and a slash denotes death. Potential probands are affected individuals, shown in red.
An Example Based on Probands

Some simple assumptions:

- Onset is late enough that the family is complete before the first case. Hence when \( \lambda_{13}(t) > 0 \), sibship size \( c \) is known and part of \( \mathcal{F}_t \).
- One parent is known to be a mutation carrier before the siblings are at risk.
- No genetic testing and no censoring (not even death).
- There are \( a(t) \) cases at time \( t^- \).
- There are \( R_t \) siblings at risk at time \( t^- \).
- Cases become probands with probability \( \pi \).
- Ascertainment occurs if there is at least one proband at time \( T \).
An Example Based on Probands

$D_{13}^i(t)$ is:

\[
 D_{13}^i(t) = \frac{(1 - (1 - \pi)^{a(t)+1}) + (1 - \pi)^{a(t)+1} \left( \prod_{j \in \mathcal{R}_t - \{i\}} [(1 - \pi) + \pi \exp (- \int_t^T \lambda_{13}^j(s) ds)] \right)}{(1 - (1 - \pi)^{a(t)}) + (1 - \pi)^{a(t)} \left( 1 - \prod_{j \in \mathcal{R}_t} [(1 - \pi) + \pi \exp (- \int_t^T \lambda_{13}^j(s) ds)] \right)}
\]

\[
 = 1 - (1 - \pi)^{a(t)+1} \left( \prod_{j \in \mathcal{R}_t - \{i\}} [(1 - \pi) + \pi \exp (- \int_t^T \lambda_{13}^j(s) ds)] \right)
\]

\[
 = \frac{1 - (1 - \pi)^{a(t)+1}}{1 - (1 - \pi)^{a(t)} \left( \prod_{j \in \mathcal{R}_t} [(1 - \pi) + \pi \exp (- \int_t^T \lambda_{13}^j(s) ds)] \right)}.
\]
An Example Based on Probands

Under complete ascertainment, $D_{13}^i = 1$ if $a(t) > 0$ and if $a(t) = 0$

$$D_{13}^i(t) = \frac{1}{1 - \prod_{j \in \mathcal{R}_t} \exp \left( - \int_t^T \lambda_{13}^j(s) ds \right)}$$

Under single ascertainment

$$D_{13}^i(t) = \frac{a(t) + 1 + \sum_{j \in \mathcal{R}_t - \{i\}} \left[ 1 - \exp \left( - \int_t^T \lambda_{13}^j(s) ds \right) \right]}{a(t) + \sum_{j \in \mathcal{R}_t} \left[ 1 - \exp \left( - \int_t^T \lambda_{13}^j(s) ds \right) \right]}.$$
Adding Censoring

Now include censoring. Results as before but in $D_{13}^{i}(t)$ replace:

$$\exp\left(-\int_{t}^{T} \lambda_{13}^{i}(s)ds\right)$$

with

$$\exp\left(-\int_{t}^{T} (\lambda_{13}^{i}(s) + \lambda_{14}^{i}(s))ds\right) + \int_{t}^{T} \exp\left(-\int_{t}^{s} (\lambda_{13}^{i}(r) + \lambda_{14}^{i}(r))dr\right) \lambda_{14}^{i}(s)ds.$$

It is obvious that $\lambda_{14}^{i}(t)$, the hazard of censoring, does not factorize out of the likelihood. This confirms earlier results by Espinosa & Macdonald (2007) who used a non-parametric estimator.

Adding genetic testing leads to the same conclusion.
Conclusions

- Elston’s likelihood can be derived in a survival model allowing for presymptomatic genetic testing. The key is formulating a marked point process model of an entire pedigree.

- Quite strong assumptions about the irrelevence of the genetic model to the decision to be tested are needed — arguably unrealistic.

- We can extend certain well-known results in a classical sibship model to the setting of survival models.

- Under retrospective ascertainment the likelihood does not factorize and terms in the hazard of disease onset alone cannot be extracted. The problem is indeed intractable.

- Actuaries using onset rates in applications should assume significant unquantifiable ascertainment bias. Example: reduce published rates by 25 – 75%.


