EMPIRIC RISK ESTIMATION IN ALZHEIMER DISEASE

By

MARK EDWARD IRWIN

B.Sc., The University of British Columbia, 1986

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

September 1989

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 September 22, 1989

ABSTRACT

Alzheimer disease is believed to be the most common cause of dementia. The main cause is presently unknown, with genetic and environmental factors suggested. It appears that 10-15% of Alzheimer disease is due to an autosomal dominant gene and it has been hypothesized that this is the cause for all Alzheimer's. Alzheimer's variable age of onset makes it more difficult to determine the validity of this and other genetic models. Empiric risk estimates for Alzheimer disease in relatives can used to test the plausibility of various genetic models.

Three types of procedures for estimating the risk of Alzheimer disease are discussed. Three nonparametric, product-limit type estimators (Kaplan-Meier, Life-table, Weinberg) for age-specific risks are discussed first. Then three estimators for lifetime risk of disease using a predetermined weight function believed to approximate the true age of onset distribution (Strömgren, Modified Strömgren, maximum likelihood) are compared. Finally a maximum likelihood procedure to estimate lifetime risk and the age of onset distribution is presented. The properties of these estimators are discussed using a data set from the Alzheimer Clinic, University Hospital - U.B.C. Site. In addition, the results of a Monte-Carlo study of the maximum likelihood procedure for estimating the lifetime risk and age of onset distribution are discussed.

The most useful of these estimators appear to be the Kaplan-Meier and the life-table estimators for age-specific risks and the maximum likelihood procedure for estimating lifetime risk and the age of onset distribution. The Weinberg estimator appears to be biased and the fixed age of onset estimators for lifetime risk appear to be too dependent on the choice of the age of onset distribution to be useful in general.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Mrs. Jean Turnbull for her help in locating the difficult to find references, Dr. Nancy Heckman for her careful reading of this thesis, and Dr. Patricia Baird for her advice and encouragement. In particular I would also like to thank Dr. John Petkau for supervising my thesis project for the past one and a half years. Finally, I would like to thank Dr. Dessa Sadovnick allowing me to use the Alzheimer data discussed in this thesis and for her advice and encouragement during the three years I worked with her.

1 INTRODUCTION

1.1 Background

Alzheimer disease (AD) is a condition clinically characterized by dementia (organic loss of cognitive function) and is often accompanied by major personality changes. It is believed to be the most common cause of dementia, accounting for 50-65% of all patients with this diagnosis (Katzman, 1976; Marsden, 1978). AD has a variable age of onset, ranging from ages 35 to 90, with the majority of people becoming affected in their 70's. The main cause of AD is presently unknown, with genetic and environmental factors hypothesized. It is believed that 10-15% of cases represent Familial Alzheimer disease (FAD), a genetic form of the disease (Friedland, 1988). These families exhibit autosomal dominant inheritance, with each child of an affected person having a 50% risk of inheriting the gene causing the disease and becoming affected themselves assuming they life long enough to reach their age of onset. A pedigree of one family appearing to represent FAD is shown in Figure 1.1 (Sadovnick et al., 1988). An explanation of the pedigree symbols is in Figure 1.2. This family is atypical, having an extremely low age of onset. It should be noted that having multiple affected members in a family does not imply that the family represents the genetic form of the disease, a "sporadic" or non-genetic form of the disease could also account for this situation. The FAD and "sporadic" forms of the disease cannot be differentiated with respect to clinical, pathological, and biochemical factors. In a few families with early onset of dementia, DNA markers have been mapped to chromosome 21 (St. George-Hyslop et al., 1987; Marx, 1988). Genetic heterogeneity in AD has been suggested by the failure of some groups to show linkage to chromosome 21 in FAD pedigrees (Schellenberg et al, 1988; Pericak-Vance et al. 1988). It has been speculated that there is no "sporadic" form of the disease, with these cases representing age-reduced penetrance of an autosomal dominant gene (Editorial, 1986).

Recent studies have suggested that the rates of AD are consistent with an autosomal dominant trait with complete penetrance by some very late age. Breitner and Folstein

(1984), Breitner et al. (1988), Martin et al. (1988), and Zubenko et al. (1988) have found risks for AD in first degree relatives approaching 50% by approximately age 90. These findings have not been consistently found, with Sadovnick et al. (1989) and Farrer et al. (1989) reporting much lower risks.

The purpose of this thesis is to investigate methods for calculating empirical risks for dementia in first-degree relatives (parents and siblings) of people with AD. These risk estimates serve two purposes. Firstly, they are useful for counselling, allowing people make better informed decisions about careers or whether to have children, for example. If someone knows that they have a 50% risk of having Alzheimer's by age 40, as in the MM family of Figure 1.1, they may decide to live their life differently than if they have risks of 10% by age 75 and 25% by age 90. Secondly, the risk estimates can be used to test the plausibility of various disease models, in particular genetic models. Of course, obtaining risk estimates consistent with an hypothesized model does not prove that the model is correct; it only provides supporting evidence.

1.2 Data Sets Investigated

The first data set investigated was collected at the Alzheimer Clinic, University Hospital - U.B.C. Site. The Clinic's multidisciplinary team consists of an internist/geriatrician, a psychiatrist, a neuropsychologist, a social worker, a geneticist, and a clinical fellow in Neurology. All patients are assessed by all members of the clinic team and are given a diagnosis according to NINCDS-ADRDA standards (McKhann et al., 1984). Risks will be calculated for relatives of patients with probable or definite AD. For a diagnosis of probable AD, dementia must be established by clinical and neuropsychological examination. There must be evidence of deficits in two or more areas of cognition, progressive worsening of memory and other cognitive functions. Also there should be no disturbance of consciousness, and no systemic disorders or other brain diseases that could account for the deficits. If in addition to the typical clinical findings, histopathological evidence from either a biopsy or autopsy consistent with AD is obtained, a definite, or autopsy

confirmed, diagnosis can be given. The pathological "hallmarks" of AD include neurofibrillary tangles, amyloid plaques, congophillic angiopathy and granulovascular change. Longitudinal studies of patients with a diagnosis of probable AD have shown that over 85% of cases have neuropathological findings consistent with definite AD (Joachim et al., 1988; Tierney et al., 1988). The diagnoses for the patients seen from January, 1985 to August 1988, the study period, are shown in Table 1.1.

Table 1.1: Diagnosis of Clinic Patients After Evaluation

Clinic Diagnosis	Number	Percentage of Total
Demented, Alzheimer's Unlikely	27	6.1
Demented, Possible Alzheimer's	90	20.2
Demented, Probable Alzheimer's	141	31.6
Definite Alzheimer's	10	2.2
Not Demented	108	24.2
Diagnosis Pending*	70	15.7
Total	446	100.0

* This category consists of patients requiring future follow-up prior to assigning a diagnosis as well as those still in the process of the assessment.

All patients referred to the clinic have, as part of their overall assessment, a detailed family history taken by a geneticist. The family history method relies on knowledgeable informants to provide the information on the relatives of the clinic patient. While the family history method has been shown to slightly underestimate the number of affected relatives when compared to the family study method in which all family members are directly assessed, the errors can be reduced by the use of multiple informants. Whenever possible, multiple informants are used, and to date over half of the families have had at least two informants. The preferred co-informants are spouses and siblings rather than the children of the clinic patients as the former tend to be more informative about older relatives. To increase the

accuracy of the medical information on the relatives, medical/autopsy records are obtained where possible. These records are evaluated by the appropriate members of the clinic team.

This method of collecting data avoids many of the biases inherent in studies in which families are ascertained through genetics clinics and solicitation of volunteers, two methods which tend to result in the over-representation of familial cases. Incorporating genetic evaluation into a specialized medical clinic has been done successfully in the past for Multiple Sclerosis, another adult-onset disease in which genetic factors appear important in the disease's etiology, but where the genetic mechanism is not clear (Sadovnick and Baird, 1988).

As only some AD may be due to a genetic trait, it is felt that for research purposes, strong criteria are needed for FAD. Of course this should be relaxed for counselling purposes as it is recognized that the following criteria only identify a very restricted group as FAD. In this study, families must satisfy four conditions to be considered as FAD as described in Sadovnick et al. (1989).

1) A detailed family history must be available must be available for at least the index case's (patient's) generation and the previous (parental) generation;
2) Good clinical documentation of dementia in relatives, preferably from at least two separate sibships within the family must be available; and there must be no other plausible explanation for the dementia such as strokes, alcoholism, head injury, etc.;
3) Neuropathological documentation of Alzheimer disease must be available for at least one member of the family, but preferably for two or more;
4) Accurate information on ages of death and/or present ages of relatives must be available so that it is possible to assess the "significance" of being clinically unaffected.

For analysis there are 825 parents and siblings of 151 consecutive, unrelated patients with probable or definite AD. Four criteria were used to determine what families would be

included in the analysis and which relatives would be classified as affected. The first three categories were used by Sadovnick et al. (1989)

a) **Stringent without FAD**: In this group, relatives were coded as "affected" only if good clinical and/or autopsy records could be obtained and Alzheimer disease seemed the almost certain diagnosis; FAD families were excluded since their inclusion could confound the results if autosomal dominant inheritance does not account for all Alzheimer disease.

b) **Stringent with FAD**: The criteria as described in (a), but FAD families are included. If all Alzheimer disease is in reality due to autosomal dominant genes, such families should be included in the analysis.

c) **Relaxed**: This includes all cases in category (b), as well as those relatives for whom the only documentation of dementia is based on the descriptions by family informants, but the descriptions do suggest dementia of unknown etiology. In particular, other causes such as strokes and cardiovascular problems have been eliminated.

d) **FAD Only**: Only members of families which have been classified as FAD according to the above rules are included in the data set. Relatives are considered affected under the stringent criteria used in (a) and (b).

A second data set involving a group of manic patients admitted to Renard Hospital, the psychiatric section of the Washington University School of Medicine, in St. Louis, between July, 1964 and June, 1965, and between January and May, 1967. The data set is described by Winokur et al. (1969). Risks for an affective disorder (mania, depression, and manic depression) will be calculated for 143 siblings of 54 manic patients. This data set is included to illustrate the properties of some of the analytic techniques.

1.3 Overview of Thesis

Three methods of estimating risks will be proposed and discussed. In Chapter 2, product-limit estimates for the probability of being affected by any given age are proposed. These are based on the non-parametric methods of survival analysis for estimating a

distribution function in the presence of censored observations. Three parametric procedures for estimating the lifetime risk for disease using a fixed predetermined approximation to the true age of onset distribution are discussed in Chapter 3. The first two are extensions of the sample proportion to estimate a binomial proportion and the third is a maximum likelihood procedure. An extension of this maximum likelihood procedure allowing the estimation of lifetime risk and the age of onset distribution is discussed in Chapter 4. This procedure also can be used to generate age-specific risk estimates similar to those calculated by the product-limit method. The results of a Monte-Carlo study investigating the properties of the extended maximum likelihood procedure are discussed in Chapter 5. In Chapter 6, the different estimation procedures are compared, and the implications the Alzheimer risk estimates have for the model that all AD is due an autosomal dominant trait are discussed.

Figure 1.1: Sample FAD Family

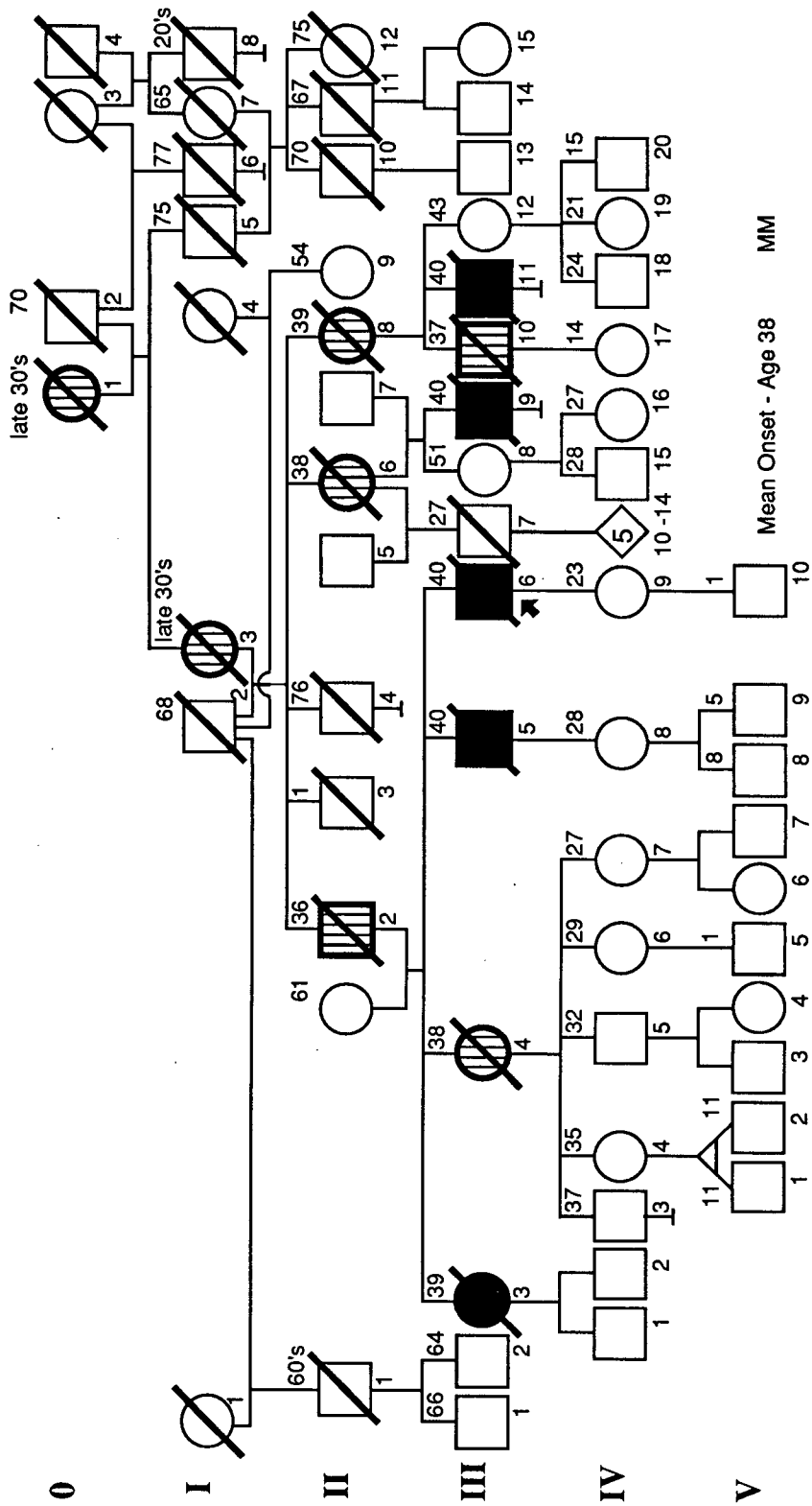
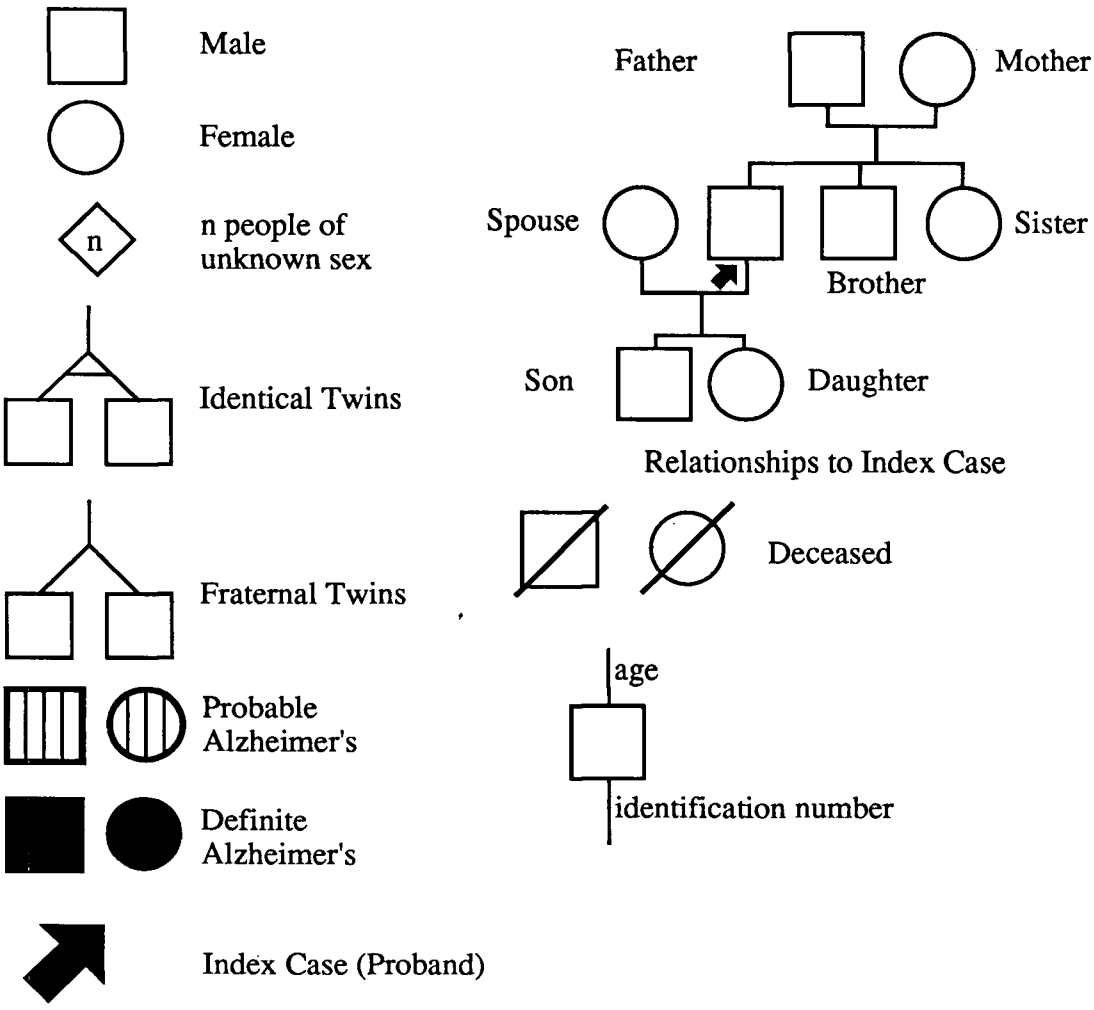


Figure 1.2: Pedigree Symbols



2 Product-Limit Estimation of Age-Specific Risks

2.1 Background

A number of groups have used product-limit type estimators to calculate age-specific risks for Alzheimer disease (Chase et al, 1983, Breitner et al, 1988, Sadovnick et al, 1989, Huff et al, 1988) and for psychiatric conditions (Slater and Cowie, 1971, Thompson and Weissman, 1981). This method has the advantage that few assumptions about the age of onset distribution need to be made. The one disadvantage of this method is that the lifetime risk cannot be estimated without making assumptions about an upper bound on the age of onset.

For these nonparametric methods to be useful, complete information about ages of family members, in particular, ages of onset for affected relatives is needed. Inaccurate determination of age of onset has been shown to lead to inconsistencies in estimation (Breitner and Magruder-Habib, 1989). Problems can also occur if the criteria used for classification as affected don't match the criteria for age of onset. Also if an affected relative's age of onset is unknown, it is not clear how this person should be dealt with. One possible solution is to consider the person unaffected at the highest age where this is clearly the case.

2.2 Model and Estimators

Assume that the probability of being affected at any given age is the same for all relatives in the sample and the outcome for each is independent of the others. Then divide the time axis into k+1 intervals $I_j = [a_{j-1}, a_j)$, $j = 1,\ldots,k+1$ where T is the largest age observed and $0 = a_0 < a_1 < \ldots < a_k = T < a_{k+1} = \infty$. For a randomly chosen relative in the sample let:

$$p_j = P[\text{does not become affected in interval } I_j \mid \text{unaffected at age } a_{j-1}]$$

$$P_j = \prod_{i=1}^{j} p_i = P[\text{unaffected at age } a_j], \text{ with } P_0 = 1 \qquad (1)$$

$$R_j = 1 - P_j = P[\text{affected by age } a_j]. \qquad (2)$$

Then collect the data into the form:

d_j = number of people with onset in interval I_j

w_j = number of people withdrawn due to censoring in interval I_j

N_j = number of people reaching age a_{j-1}.

The product-limit estimates are based on choosing estimators for p_j. Three estimators which have been proposed are:

$$p_j^W = \frac{N_j - \frac{1}{2}(w_j + d_j) - d_j}{N_j - \frac{1}{2}(w_j + d_j)} \qquad \text{Weinberg}$$

$$p_j^{LT} = \frac{N_j - \frac{1}{2}w_j - d_j}{N_j - \frac{1}{2}w_j} \qquad \text{Life-Table}$$

$$p_j^{KM} = \frac{N_j - d_j}{N_j} \qquad \text{Kaplan-Meier (1958).}$$

These estimates lead to the age-specific escape probability estimates, P_j^W, P_j^{LT} and P_j^{KM}, which are calculated by substituting the appropriate estimate for p_i into equation (1). Then the age-specific risk estimates, R_j^W, R_j^{LT} and R_j^{KM}, are calculated by using the appropriate estimate for P_j in equation (2).

These three different estimators result from different assumptions about the censoring and onset patterns in each age interval. The Weinberg estimate is based on the assumption that onsets and withdrawals occur uniformly within each interval. The actuarial life-table method assumes only withdrawals occur uniformly within an age interval, with no assumptions made about where in the interval the onsets occur. The Kaplan-Meier procedure makes the assumption that all withdrawals occur after all onsets within each age interval.

An estimate for the variance of P_j^W and R_j^W (Slater and Cowie, 1971) is:

$$\mathrm{Var}\left[P_j^W\right] = \mathrm{Var}\left[R_j^W\right] = \left(P_j^W\right)^2 \sum_{i=1}^{j} \frac{1 - p_i^W}{\left(N_i - \frac{1}{2}(w_i + d_i)\right) p_i^W}$$

$$= \left(P_j^W\right)^2 \sum_{i=1}^{j} \frac{d_i}{\left(N_i - \frac{1}{2}(w_i + d_i)\right)\left(N_i - \frac{1}{2}(w_i + d_i) - d_i\right)} .$$

The estimate for the variance of P_j^{LT} and R_j^{LT} as shown by Greenwood (1926) is:

$$\mathrm{Var}\left[P_j^{LT}\right] = \mathrm{Var}\left[R_j^{LT}\right] = \left(P_j^{LT}\right)^2 \sum_{i=1}^{j} \frac{1 - p_i^{LT}}{\left(N_i - \frac{1}{2}w_i\right) p_i^{LT}}$$

$$= \left(P_j^{LT}\right)^2 \sum_{i=1}^{j} \frac{d_i}{\left(N_i - \frac{1}{2}w_i\right)\left(N_i - \frac{1}{2}w_i - d_i\right)} .$$

The similar estimate for the variance of P_j^{KM} and R_j^{KM} as shown by Kaplan and Meier (1958) is:

$$\mathrm{Var}\left[P_j^{KM}\right] = \mathrm{Var}\left[R_j^{KM}\right] = \left(P_j^{KM}\right)^2 \sum_{i=1}^{j} \frac{1 - p_i^{KM}}{N_i\, p_i^{KM}}$$

$$= \left(P_j^{KM}\right)^2 \sum_{i=1}^{j} \frac{d_i}{N_i\left(N_i - d_i\right)} .$$

These variance estimates can be used to calculate confidence intervals of the escape and risk probabilities at any given age. One possible choice (Lawless, 1982) for an approximate 100 $(1 - \alpha)$% confidence interval for P_j using the asymptotic normality of the above estimates is:

$$P_j^* \pm z_{\alpha/2} \sqrt{\mathrm{Var}\left[P_j^*\right]}$$

where P_j^* is one of three discussed estimates and $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the normal distribution. Another option (Lawless, 1982) is to apply a transform ψ so that the

distribution of $\psi(P_j^*)$ is closer to a normal than the distribution of P_j^*. One good choice for ψ (Anscombe, 1964) is:

$$\psi(P) = \int_0^P t^{-1/3}(1-t)^{-1/3}\,dt .$$

Another choice which is almost as good (Lawless, 1982) is the logistic transform

$$\psi(P) = \log\left(\frac{P}{1-P}\right).$$

The variance of $\psi(P_j^*)$ can be estimated using the delta method by:

$$S^2_{\widehat{\psi}} = \left[\psi'\left(P_j^*\right)\right]^2 \mathrm{Var}\left[P_j^*\right].$$

If ψ_L and ψ_U are defined as:

$$\psi_L = \psi\left(P_j^*\right) - z_{\alpha/2}S_{\widehat{\psi}}\,,\quad \psi_U = \psi\left(P_j^*\right) + z_{\alpha/2}S_{\widehat{\psi}}\,,$$

then a 100 $(1-\alpha)$% confidence interval for P_j is $(\psi^{-1}(\psi_L)\,,\,\psi^{-1}(\psi_U))$ where ψ^{-1} is the inverse function of ψ. Confidence intervals for the age-specific risks can be derived using the relationship between P_j and R_j.

2.3 Properties of Estimators

The three estimators proposed in section 2.2 satisfy the following orderings:

i) $p_j^W \le p_j^{LT} \le p_j^{KM}$

ii) $P_j^W \le P_j^{LT} \le P_j^{KM}$

iii) $R_j^W \ge R_j^{LT} \ge R_j^{KM}$

The relationship between the estimators of p_j can easily be seen by examining $p_j^{LT} - p_j^W$ and $p_j^{KM} - p_j^{LT}$.

$$p_j^{LT} - p_j^W = \frac{\frac{1}{2}d_j^2}{\left(N_j - \frac{1}{2}w_j\right)\left(N_j - \frac{1}{2}(w_j + d_j)\right)}$$

$$= p_j^{LT}\frac{\frac{1}{2}d_j^2}{\left(N_j - \frac{1}{2}w_j - d_j\right)\left(N_j - \frac{1}{2}(w_j + d_j)\right)} \ge 0 .$$

$$p_j^{KM} - p_j^{LT} = \frac{\frac{1}{2}d_j w_j}{N_j\left(N_j - \frac{1}{2}w_j\right)}$$

$$= p_j^{KM} \frac{\frac{1}{2}d_j w_j}{\left(N_j - d_j\right)\left(N_j - \frac{1}{2}w_j\right)} \geq 0 .$$

The relationships between the estimators P_j and R_j are then obvious corollaries.

In any interval I_j, $p_j^{LT} = p_j^{KM}$ if either $w_j = 0$ or $d_j = 0$. This implies the well known property that if all the withdrawals occur in intervals after all the onsets, the life-table and the Kaplan-Meier estimates will be the same.

A small simulation study by Chase et al. (1983) suggested that the Kaplan-Meier estimator is approximately unbiased. This is not surprising as Kaplan and Meier (1958) showed that P_j^{KM} is a consistent estimator (and therefore so is R_j^{KM}) when some reasonable assumptions are made. Chase et al. also stated that the life-table estimator appears to be approximately unbiased. This statement is mildly surprising since it is known that this estimator is not consistent (Lawless, 1982). Finally, Chase et al indicated that the Weinberg risk estimator appears to have a positive bias. This is to be expected due to the above ordering of the estimators. The size of the bias does not appear large in their trials, however they do not give any indication as to the size of the bias in general.

2.4 Results

All three estimation procedures were used to calculate age-specific risks in Alzheimer disease under the four diagnostic criteria discussed in Chapter 1. The estimated risks are shown in Tables 2.1 - 2.4 , with plots of the risks shown in Figures 2.1 - 2.4. As can be clearly seen in the figures, the difference between the Weinberg and life-table estimates is much smaller than the difference between the life-table and the Kaplan-Meier estimates, with the size of the deviations increasing with age. This is due to the relatively heavy censoring in this data set. This heavy censoring is to be expected if the risk for Alzheimer's was 50% by age 90, on average well over half the members of the sample would have

censored observations. However, all three estimators show similar risk curves under the four diagnostic criteria.

In particular, except for the FAD only criteria, the risk estimates are not consistent with the 50% risks by age 90 found by other researchers. This suggests that not all Alzheimer's is due to an autosomal dominant trait with complete penetrance by age 90. The risk for dementia by age 90 under the four criteria and the three estimators is shown in Table 2.5. However, this method calculates risks to certain ages, not lifetime risk. The latter is what is needed to make better statements about the plausibility of the autosomal dominant model.

Table 2.1: Age-Specific Risks Under Stringent without FAD Criteria

Age	Total	Affected	Withdrawn	Weinberg	Life-Table	Kaplan-Meier
0	766	0	13	0.000	0.000	0.000
1	753	0	3	0.000	0.000	0.000
2	750	0	5	0.000	0.000	0.000
3	745	0	1	0.000	0.000	0.000
4	744	0	2	0.000	0.000	0.000
5	742	0	1	0.000	0.000	0.000
6	741	0	2	0.000	0.000	0.000
7	739	0	0	0.000	0.000	0.000
8	739	0	1	0.000	0.000	0.000
9	738	0	0	0.000	0.000	0.000
10	738	0	0	0.000	0.000	0.000
11	738	0	0	0.000	0.000	0.000
12	738	0	1	0.000	0.000	0.000
13	737	0	0	0.000	0.000	0.000
14	737	0	0	0.000	0.000	0.000
15	737	0	0	0.000	0.000	0.000
16	737	0	0	0.000	0.000	0.000
17	737	0	0	0.000	0.000	0.000
18	737	0	5	0.000	0.000	0.000
19	732	0	1	0.000	0.000	0.000
20	731	0	9	0.000	0.000	0.000
21	722	0	8	0.000	0.000	0.000
22	714	0	3	0.000	0.000	0.000
23	711	0	1	0.000	0.000	0.000
24	710	0	1	0.000	0.000	0.000
25	709	0	5	0.000	0.000	0.000
26	704	0	0	0.000	0.000	0.000
27	704	0	1	0.000	0.000	0.000
28	703	0	3	0.000	0.000	0.000
29	700	0	3	0.000	0.000	0.000
30	697	0	6	0.000	0.000	0.000
31	691	0	2	0.000	0.000	0.000
32	689	0	0	0.000	0.000	0.000
33	689	0	2	0.000	0.000	0.000
34	687	0	0	0.000	0.000	0.000
35	687	0	10	0.000	0.000	0.000
36	677	0	0	0.000	0.000	0.000
37	677	0	0	0.000	0.000	0.000
38	677	0	3	0.000	0.000	0.000
39	674	0	0	0.000	0.000	0.000
40	674	0	9	0.000	0.000	0.000
41	665	0	1	0.000	0.000	0.000
42	664	0	3	0.000	0.000	0.000
43	661	0	2	0.000	0.000	0.000
44	659	0	0	0.000	0.000	0.000
45	659	0	8	0.000	0.000	0.000
46	651	0	1	0.000	0.000	0.000

47	650	0	2	0.000	0.000	0.000
48	648	0	5	0.000	0.000	0.000
49	643	0	4	0.000	0.000	0.000
50	639	1	6	0.000	0.000	0.000
51	632	0	1	0.002	0.002	0.002
52	631	1	4	0.002	0.002	0.002
53	626	0	5	0.003	0.003	0.003
54	621	0	3	0.003	0.003	0.003
55	618	0	17	0.003	0.003	0.003
56	601	0	9	0.003	0.003	0.003
57	592	0	7	0.003	0.003	0.003
58	585	0	14	0.003	0.003	0.003
59	571	0	5	0.003	0.003	0.003
60	566	1	24	0.003	0.003	0.003
61	541	1	5	0.005	0.005	0.005
62	535	1	12	0.007	0.007	0.007
63	522	1	21	0.009	0.009	0.009
64	500	1	13	0.011	0.011	0.011
65	486	1	32	0.013	0.013	0.012
66	453	0	15	0.015	0.015	0.015
67	438	1	21	0.015	0.015	0.015
68	416	0	21	0.017	0.017	0.017
69	395	2	15	0.017	0.017	0.017
70	378	1	43	0.022	0.022	0.022
71	334	1	10	0.025	0.025	0.024
72	323	1	28	0.028	0.028	0.027
73	294	2	20	0.031	0.031	0.030
74	272	3	13	0.038	0.038	0.037
75	256	1	40	0.049	0.049	0.047
76	215	2	18	0.053	0.053	0.051
77	195	0	9	0.062	0.062	0.060
78	186	0	27	0.062	0.062	0.060
79	159	1	8	0.062	0.062	0.060
80	150	3	36	0.068	0.068	0.066
81	111	0	11	0.090	0.089	0.085
82	100	1	8	0.090	0.089	0.085
83	91	0	6	0.099	0.099	0.094
84	85	0	16	0.099	0.099	0.094
85	69	1	20	0.099	0.099	0.094
86	48	0	12	0.114	0.114	0.107
87	36	0	8	0.114	0.114	0.107
88	28	0	4	0.114	0.114	0.107
89	24	0	5	0.114	0.114	0.107
90	19	0	3	0.114	0.114	0.107
91	16	0	1	0.114	0.114	0.107
92	15	0	5	0.114	0.114	0.107
93	10	0	2	0.114	0.114	0.107
94	8	0	1	0.114	0.114	0.107
95	7	0	1	0.114	0.114	0.107
96	6	0	3	0.114	0.114	0.107
97	3	0	2	0.114	0.114	0.107
98	1	0	1	0.114	0.114	0.107

Table 2.2: Age-Specific Risks Under Stringent with FAD Criteria

Age	Total	Affected	Withdrawn	Weinberg	Life-Table	Kaplan-Meier
0	824	0	14	0.000	0.000	0.000
1	810	0	3	0.000	0.000	0.000
2	807	0	5	0.000	0.000	0.000
3	802	0	1	0.000	0.000	0.000
4	801	0	1	0.000	0.000	0.000
5	800	0	2	0.000	0.000	0.000
6	798	0	2	0.000	0.000	0.000
7	796	0	0	0.000	0.000	0.000
8	796	0	1	0.000	0.000	0.000
9	795	0	1	0.000	0.000	0.000
10	794	0	0	0.000	0.000	0.000
11	794	0	0	0.000	0.000	0.000
12	794	0	1	0.000	0.000	0.000
13	793	0	0	0.000	0.000	0.000
14	793	0	0	0.000	0.000	0.000
15	793	0	0	0.000	0.000	0.000
16	793	0	1	0.000	0.000	0.000
17	792	0	0	0.000	0.000	0.000
18	792	0	5	0.000	0.000	0.000
19	787	0	1	0.000	0.000	0.000
20	786	0	9	0.000	0.000	0.000
21	777	0	8	0.000	0.000	0.000
22	769	0	3	0.000	0.000	0.000
23	766	0	1	0.000	0.000	0.000
24	765	0	2	0.000	0.000	0.000
25	763	0	5	0.000	0.000	0.000
26	758	0	0	0.000	0.000	0.000
27	758	0	1	0.000	0.000	0.000
28	757	0	3	0.000	0.000	0.000
29	754	0	3	0.000	0.000	0.000
30	751	0	7	0.000	0.000	0.000
31	744	0	2	0.000	0.000	0.000
32	742	0	0	0.000	0.000	0.000
33	742	0	2	0.000	0.000	0.000
34	740	0	0	0.000	0.000	0.000
35	740	0	10	0.000	0.000	0.000
36	730	1	0	0.000	0.000	0.000
37	729	0	1	0.001	0.001	0.001
38	728	1	3	0.001	0.001	0.001
39	724	1	0	0.003	0.003	0.003
40	723	1	10	0.004	0.004	0.004
41	712	0	1	0.006	0.006	0.005
42	711	0	3	0.006	0.006	0.005
43	708	0	2	0.006	0.006	0.005
44	706	0	2	0.006	0.006	0.005
45	704	0	8	0.006	0.006	0.005
46	696	0	2	0.006	0.006	0.005

47	694	0	2	0.006	0.006	0.005
48	692	0	6	0.006	0.006	0.005
49	686	0	4	0.006	0.006	0.005
50	682	1	6	0.006	0.006	0.005
51	675	0	1	0.007	0.007	0.007
52	674	1	6	0.007	0.007	0.007
53	667	0	6	0.008	0.008	0.008
54	661	0	4	0.008	0.008	0.008
55	657	1	19	0.008	0.008	0.008
56	637	0	10	0.010	0.010	0.010
57	627	0	8	0.010	0.010	0.010
58	619	0	15	0.010	0.010	0.010
59	604	1	5	0.010	0.010	0.010
60	598	1	25	0.012	0.012	0.012
61	572	1	7	0.013	0.013	0.013
62	564	1	14	0.015	0.015	0.015
63	549	1	22	0.017	0.017	0.017
64	526	2	14	0.019	0.019	0.018
65	510	1	32	0.022	0.022	0.022
66	477	0	15	0.024	0.024	0.024
67	462	3	21	0.024	0.024	0.024
68	438	0	21	0.031	0.031	0.030
69	417	2	16	0.031	0.031	0.030
70	399	3	44	0.036	0.036	0.035
71	352	1	13	0.043	0.043	0.042
72	338	1	29	0.046	0.046	0.045
73	308	3	20	0.049	0.049	0.048
74	285	3	13	0.059	0.059	0.057
75	269	2	40	0.069	0.069	0.067
76	227	2	18	0.076	0.076	0.074
77	207	1	10	0.085	0.085	0.082
78	196	0	28	0.089	0.089	0.087
79	168	1	8	0.089	0.089	0.087
80	159	3	36	0.095	0.095	0.092
81	120	1	12	0.115	0.114	0.109
82	107	1	9	0.122	0.122	0.117
83	97	0	8	0.131	0.130	0.125
84	89	0	16	0.131	0.130	0.125
85	73	1	21	0.131	0.130	0.125
86	51	0	13	0.145	0.144	0.137
87	38	0	8	0.145	0.144	0.137
88	30	0	4	0.145	0.144	0.137
89	26	0	5	0.145	0.144	0.137
90	21	0	3	0.145	0.144	0.137
91	18	0	1	0.145	0.144	0.137
92	17	0	5	0.145	0.144	0.137
93	12	0	3	0.145	0.144	0.137
94	9	0	1	0.145	0.144	0.137
95	8	0	2	0.145	0.144	0.137
96	6	0	3	0.145	0.144	0.137
97	3	0	2	0.145	0.144	0.137
98	1	0	1	0.145	0.144	0.137

Table 2.3: Age-Specific Risks Under Relaxed Criteria

Age	Total	Affected	Withdrawn	Weinberg	Life-Table	Kaplan-Meier
0	825	0	14	0.000	0.000	0.000
1	811	0	3	0.000	0.000	0.000
2	808	0	5	0.000	0.000	0.000
3	803	0	1	0.000	0.000	0.000
4	802	0	1	0.000	0.000	0.000
5	801	0	2	0.000	0.000	0.000
6	799	0	2	0.000	0.000	0.000
7	797	0	0	0.000	0.000	0.000
8	797	0	1	0.000	0.000	0.000
9	796	0	1	0.000	0.000	0.000
10	795	0	0	0.000	0.000	0.000
11	795	0	0	0.000	0.000	0.000
12	795	0	1	0.000	0.000	0.000
13	794	0	0	0.000	0.000	0.000
14	794	0	0	0.000	0.000	0.000
15	794	0	0	0.000	0.000	0.000
16	794	0	1	0.000	0.000	0.000
17	793	0	0	0.000	0.000	0.000
18	793	0	5	0.000	0.000	0.000
19	788	0	1	0.000	0.000	0.000
20	787	0	9	0.000	0.000	0.000
21	778	0	8	0.000	0.000	0.000
22	770	0	3	0.000	0.000	0.000
23	767	0	1	0.000	0.000	0.000
24	766	0	2	0.000	0.000	0.000
25	764	0	5	0.000	0.000	0.000
26	759	0	0	0.000	0.000	0.000
27	759	0	1	0.000	0.000	0.000
28	758	0	3	0.000	0.000	0.000
29	755	0	3	0.000	0.000	0.000
30	752	0	7	0.000	0.000	0.000
31	745	0	2	0.000	0.000	0.000
32	743	0	0	0.000	0.000	0.000
33	743	0	2	0.000	0.000	0.000
34	741	0	0	0.000	0.000	0.000
35	741	0	10	0.000	0.000	0.000
36	731	1	0	0.000	0.000	0.000
37	730	0	1	0.001	0.001	0.001
38	729	1	3	0.001	0.001	0.001
39	725	1	0	0.003	0.003	0.003
40	724	1	10	0.004	0.004	0.004
41	713	0	1	0.006	0.006	0.005
42	712	0	3	0.006	0.006	0.005
43	709	0	2	0.006	0.006	0.005
44	707	0	2	0.006	0.006	0.005
45	705	0	8	0.006	0.006	0.005
46	697	0	2	0.006	0.006	0.005

47	695	0	2	0.006	0.006	0.005
48	693	0	6	0.006	0.006	0.005
49	687	0	4	0.006	0.006	0.005
50	683	1	6	0.006	0.006	0.005
51	676	0	1	0.007	0.007	0.007
52	675	1	6	0.007	0.007	0.007
53	668	0	6	0.008	0.008	0.008
54	662	0	4	0.008	0.008	0.008
55	658	1	19	0.008	0.008	0.008
56	638	0	10	0.010	0.010	0.010
57	628	0	8	0.010	0.010	0.010
58	620	0	15	0.010	0.010	0.010
59	605	2	5	0.010	0.010	0.010
60	598	2	25	0.013	0.013	0.013
61	571	1	7	0.017	0.017	0.016
62	563	1	14	0.018	0.018	0.018
63	548	1	22	0.020	0.020	0.020
64	525	2	13	0.022	0.022	0.022
65	510	1	32	0.026	0.026	0.025
66	477	0	15	0.028	0.028	0.027
67	462	3	21	0.028	0.028	0.027
68	438	0	21	0.034	0.034	0.034
69	417	2	16	0.034	0.034	0.034
70	399	3	44	0.039	0.039	0.038
71	352	1	13	0.047	0.047	0.046
72	338	2	29	0.049	0.049	0.048
73	307	3	20	0.055	0.055	0.054
74	284	6	13	0.065	0.065	0.063
75	265	4	40	0.085	0.085	0.083
76	221	2	18	0.100	0.100	0.097
77	201	1	10	0.109	0.108	0.105
78	190	0	24	0.113	0.113	0.109
79	166	2	8	0.113	0.113	0.109
80	156	4	35	0.124	0.124	0.120
81	117	1	12	0.150	0.149	0.143
82	104	2	9	0.158	0.157	0.150
83	93	1	8	0.175	0.174	0.166
84	84	1	16	0.184	0.183	0.175
85	67	1	18	0.195	0.194	0.185
86	48	0	12	0.209	0.208	0.197
87	36	1	8	0.209	0.208	0.197
88	27	0	3	0.234	0.232	0.220
89	24	0	5	0.234	0.232	0.220
90	19	0	3	0.234	0.232	0.220
91	16	0	0	0.234	0.232	0.220
92	16	0	5	0.234	0.232	0.220
93	11	0	3	0.234	0.232	0.220
94	8	0	0	0.234	0.232	0.220
95	8	0	2	0.234	0.232	0.220
96	6	0	3	0.234	0.232	0.220
97	3	0	2	0.234	0.232	0.220
98	1	0	1	0.234	0.232	0.220

Table 2.4: Age-Specific Risks Under FAD Only Criteria

Age	Total	Affected	Withdrawn	Weinberg	Life-Table	Kaplan-Meier
0	62	0	3	0.000	0.000	0.000
1	59	0	0	0.000	0.000	0.000
2	59	0	0	0.000	0.000	0.000
3	59	0	0	0.000	0.000	0.000
4	59	0	0	0.000	0.000	0.000
5	59	0	1	0.000	0.000	0.000
6	58	0	0	0.000	0.000	0.000
7	58	0	0	0.000	0.000	0.000
8	58	0	0	0.000	0.000	0.000
9	58	0	1	0.000	0.000	0.000
10	57	0	0	0.000	0.000	0.000
11	57	0	0	0.000	0.000	0.000
12	57	0	0	0.000	0.000	0.000
13	57	0	0	0.000	0.000	0.000
14	57	0	0	0.000	0.000	0.000
15	57	0	0	0.000	0.000	0.000
16	57	0	1	0.000	0.000	0.000
17	56	0	0	0.000	0.000	0.000
18	56	0	0	0.000	0.000	0.000
19	56	0	0	0.000	0.000	0.000
20	56	0	0	0.000	0.000	0.000
21	56	0	0	0.000	0.000	0.000
22	56	0	0	0.000	0.000	0.000
23	56	0	0	0.000	0.000	0.000
24	56	0	1	0.000	0.000	0.000
25	55	0	0	0.000	0.000	0.000
26	55	0	0	0.000	0.000	0.000
27	55	0	0	0.000	0.000	0.000
28	55	0	0	0.000	0.000	0.000
29	55	0	0	0.000	0.000	0.000
30	55	0	1	0.000	0.000	0.000
31	54	0	0	0.000	0.000	0.000
32	54	0	0	0.000	0.000	0.000
33	54	0	0	0.000	0.000	0.000
34	54	0	0	0.000	0.000	0.000
35	54	0	0	0.000	0.000	0.000
36	54	1	0	0.000	0.000	0.000
37	53	0	1	0.019	0.019	0.019
38	52	1	0	0.019	0.019	0.019
39	51	1	0	0.038	0.037	0.037
40	50	1	1	0.057	0.056	0.056
41	48	0	0	0.076	0.075	0.075
42	48	0	0	0.076	0.075	0.075
43	48	0	0	0.076	0.075	0.075
44	48	0	2	0.076	0.075	0.075
45	46	0	0	0.076	0.075	0.075
46	46	0	1	0.076	0.075	0.075

47	45	0	0	0.076	0.075	0.075
48	45	0	1	0.076	0.075	0.075
49	44	0	0	0.076	0.075	0.075
50	44	0	0	0.076	0.075	0.075
51	44	0	0	0.076	0.075	0.075
52	44	0	2	0.076	0.075	0.075
53	42	0	1	0.076	0.075	0.075
54	41	0	1	0.076	0.075	0.075
55	40	1	2	0.076	0.075	0.075
56	37	0	2	0.100	0.099	0.098
57	35	0	1	0.100	0.099	0.098
58	34	0	1	0.100	0.099	0.098
59	33	1	0	0.100	0.099	0.098
60	32	0	0	0.128	0.126	0.126
61	32	0	2	0.128	0.126	0.126
62	30	0	2	0.128	0.126	0.126
63	28	0	1	0.128	0.126	0.126
64	27	1	0	0.128	0.126	0.126
65	26	0	0	0.161	0.159	0.158
66	26	0	0	0.161	0.159	0.158
67	26	2	0	0.161	0.159	0.158
68	24	0	0	0.228	0.223	0.223
69	24	0	1	0.228	0.223	0.223
70	23	3	1	0.228	0.223	0.223
71	19	0	3	0.338	0.327	0.324
72	16	0	1	0.338	0.327	0.324
73	15	2	0	0.338	0.327	0.324
74	13	0	0	0.433	0.417	0.414
75	13	1	0	0.433	0.417	0.414
76	12	0	0	0.478	0.462	0.459
77	12	1	1	0.478	0.462	0.459
78	10	0	1	0.526	0.508	0.504
79	9	0	0	0.526	0.508	0.504
80	9	0	0	0.526	0.508	0.504
81	9	1	1	0.526	0.508	0.504
82	7	0	1	0.585	0.566	0.559
83	6	0	2	0.585	0.566	0.559
84	4	0	0	0.585	0.566	0.559
85	4	0	1	0.585	0.566	0.559
86	3	0	1	0.585	0.566	0.559
87	2	0	0	0.585	0.566	0.559
88	2	0	0	0.585	0.566	0.559
89	2	0	0	0.585	0.566	0.559
90	2	0	0	0.585	0.566	0.559
91	2	0	0	0.585	0.566	0.559
92	2	0	0	0.585	0.566	0.559
93	2	0	1	0.585	0.566	0.559
94	1	0	0	0.585	0.566	0.559
95	1	0	1	0.585	0.566	0.559
96	0	0	0	0.585	0.566	0.559

Table 2.5: Risk for Dementia at Age 90 with Standard Errors.

Criteria	Weinberg	Life-table	Kaplan-Meier
Stringent without FAD	0.114 (0.026)	0.114 (0.026)	0.107 (0.023)
Stringent with FAD	0.145 (0.026)	0.144 (0.026)	0.137 (0.024)
Relaxed	0.234 (0.039)	0.232 (0.039)	0.220 (0.036)
FAD Only	0.585 (0.103)	0.566 (0.102)	0.559 (0.101)

Figure 2.1: Age-Specific Risks Under Stringent without FAD Criteria

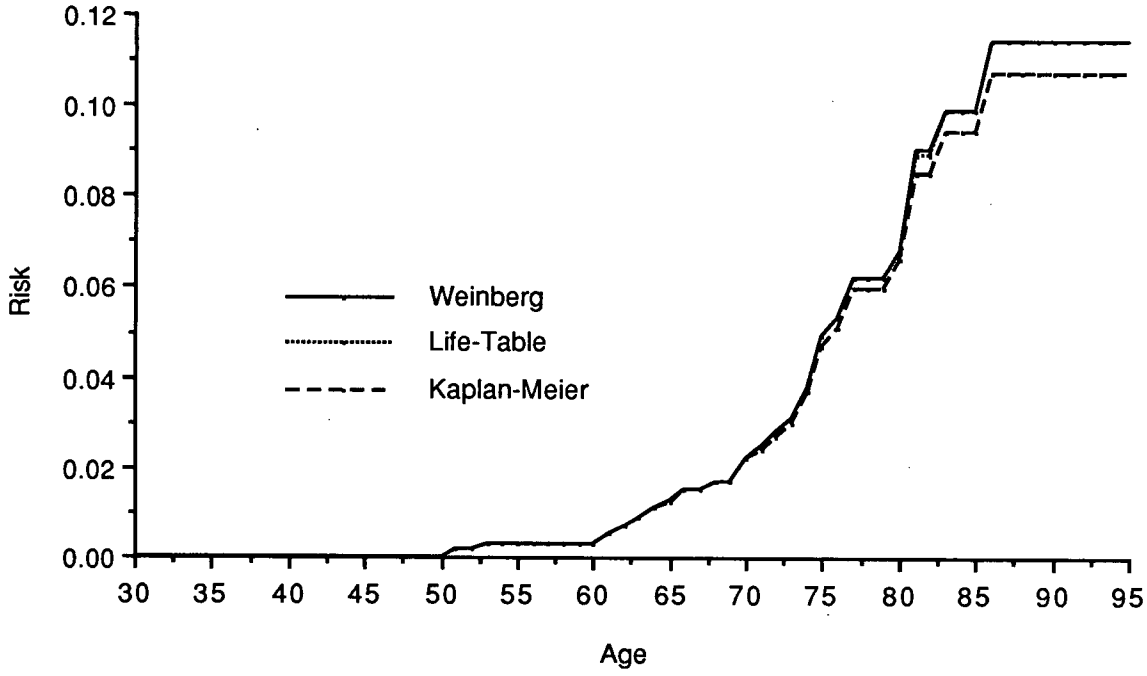


Figure 2.2: Age-Specific Risks Under Stringent with FAD Criteria

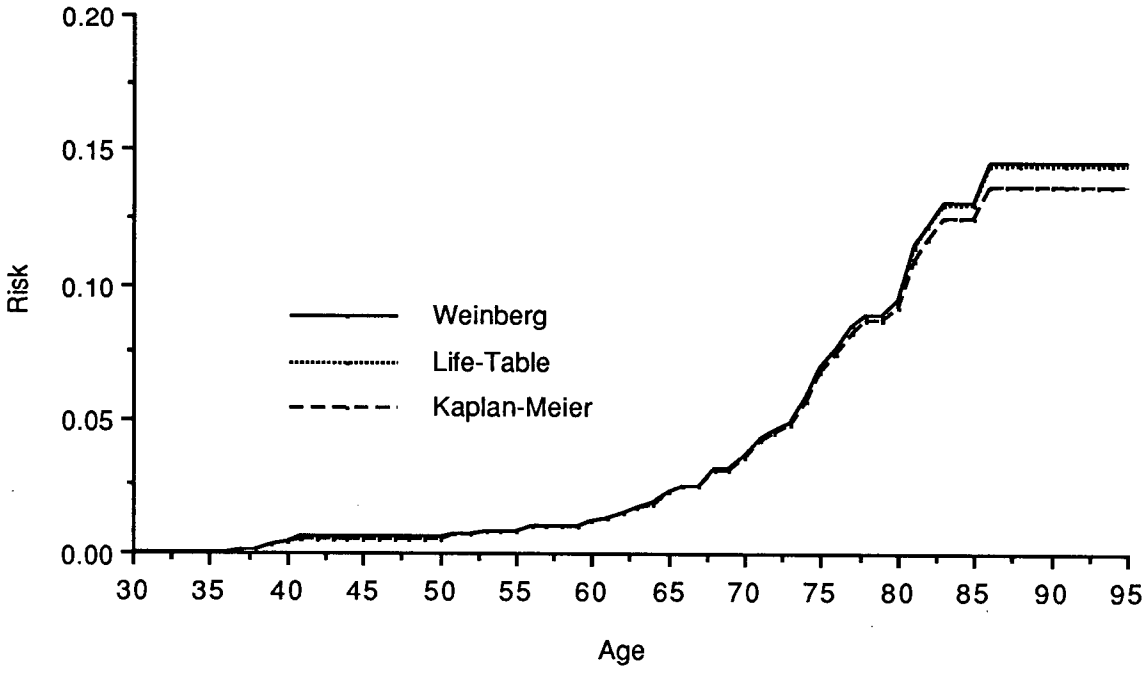


Figure 2.3: Age-Specific Risks Under Relaxed Criteria

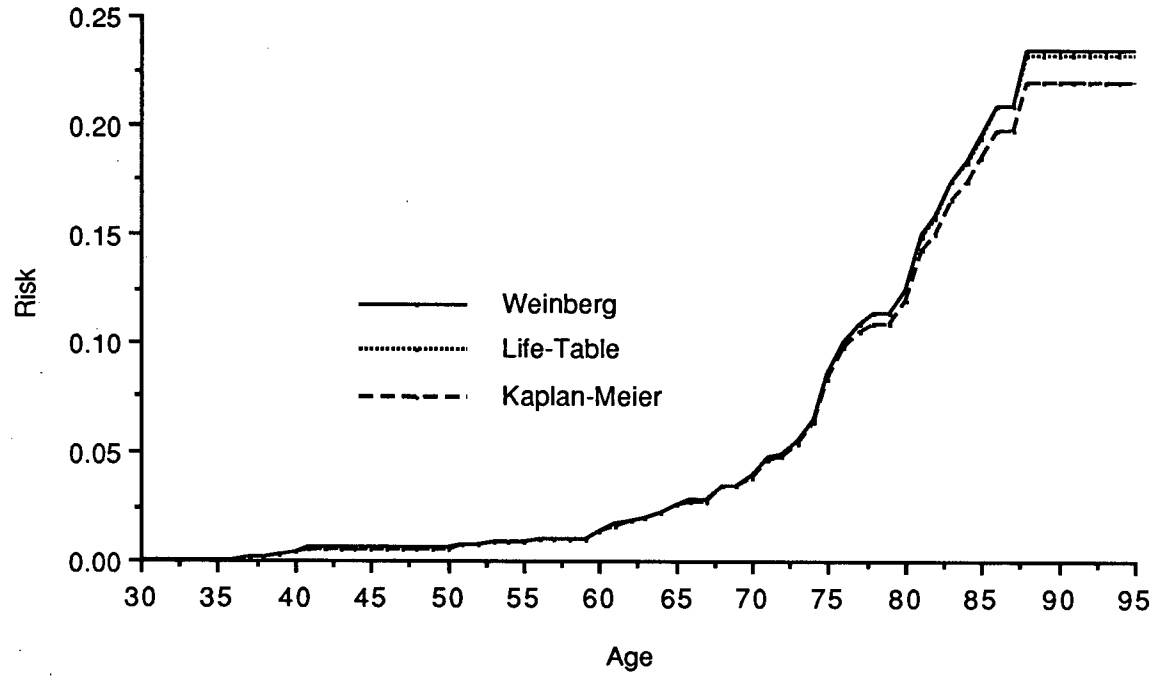
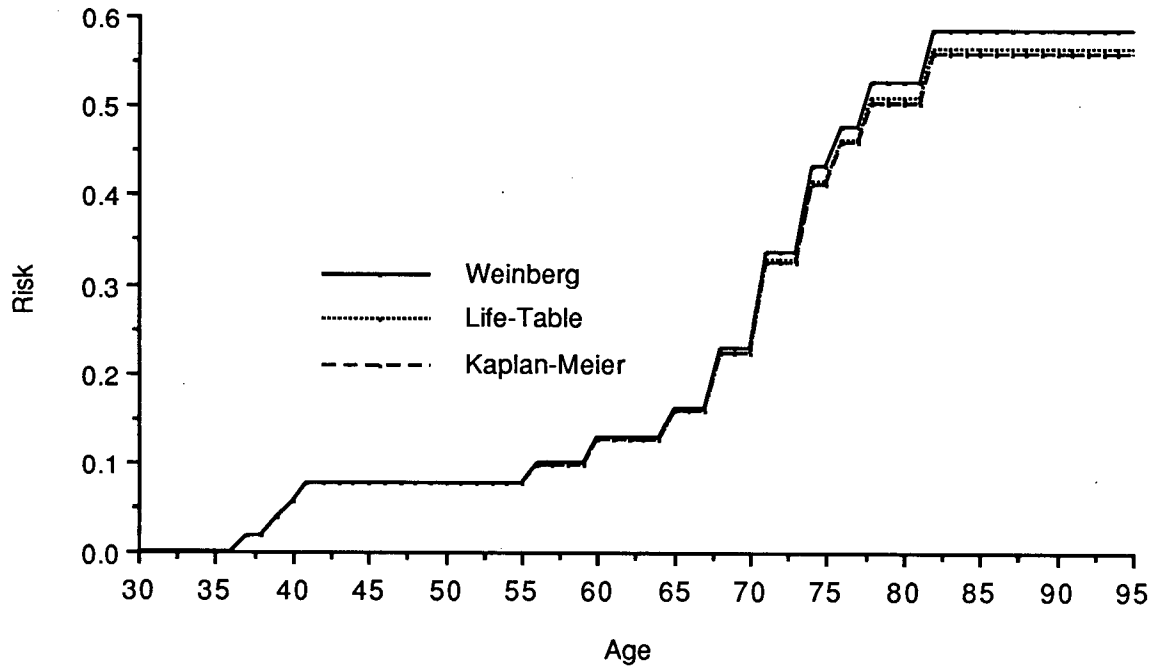


Figure 2.4: Age-Specific Risks Under FAD Only Criteria



3 Lifetime Risk Estimation Using Fixed Age of Onset Distributions

3.1 Background

Early methods for estimating the proportion of the relatives susceptible to disease, or their lifetime risk, p, used a fixed, predetermined weight function w(t), believed to approximate the true age of onset distribution F(t) = P[affected by age t | susceptible] in the relatives. The age of onset distribution gives a measure of risk experienced, assuming the relative would get the disease if they lived long enough. Since some unaffected relatives could still be at risk for disease, using this weight function should give a better estimate of lifetime risk than the biased sample proportion of affected relatives.

3.2 Model and Estimators

Similarly to the product-limit procedure, assume that the lifetime risk and the age of onset distribution are the same for each relative in the sample and that the status for each relative is independent of the rest. The data required for relative i, i=1,...,n in the sample is the pair (x_i,t_i) where

$$x_i = \begin{cases} 0 & \text{if relative i is unaffected} \\ 1 & \text{if relative i is affected} \end{cases}$$

$$t_i = \text{age of observation for relative i} = \begin{cases} \text{age at death} & \text{if relative i has died} \\ \text{current age} & \text{if relative i is alive} \end{cases}$$

Denote the true and approximate conditional risks for relative i as $f_i = F(t_i)$ and $w_i = w(t_i)$. If the random variable X_i denotes the status of relative i at age t_i, $E[X_i] = pF(t_i) = pf_i$.

Some general weight functions used in the past are:

1) $$w(t) = \begin{cases} 0 & t < a_1 \\ \frac{1}{2} & a_1 \le t \le a_2 \\ 1 & t > a_2 \end{cases}$$ Weinberg (1925)

2) $w(t) = \begin{cases} 0 & t < \bar{a} \\ 1 & t \geq \bar{a} \end{cases}$ Larsson & Strögren (1954)

3) $w(t) = \begin{cases} 0 & t < a_1 \\ \frac{t - a_1}{a_2 - a_1} & a_1 \leq t \leq a_2 \\ 1 & t > a_2 \end{cases}$ Schulz (1937)

where $[a_1,a_2]$ is the age range of susceptibility and $\bar{a}$ is the mean age of onset.

4) Strömgren (1935) recommended that a previously observed age of onset distribution be used. Similar to this is the use of the empirical distribution function of the ages of onset of the index cases as used by Winokur et al. (1969).

A valid weight function w(t) is one that satisfies the following conditions:

1) $w(t):[0,\infty) \rightarrow [0,1]$,

2) w(t) is non-decreasing,

3) $\lim_{t\rightarrow\infty} w(t) = 1$

While the condition $w(0) = 0$ is not necessary in theory, it is usually appropriate. The situation where $w(0) > 0$ implies that the condition can be present at birth.

The following three estimates for lifetime risk have been proposed.

3.2.1 Original Strömgren (1935):

$$p^* = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} w_i}$$

This estimator is an extension of the sample proportion. This estimator has the undesirable property that it is possible for $p^* > 1$. However in most situations the probability of this happening should be extremely low.

The bias of p^* is easily calculated:

$$\text{Bias}(p^*) = E[p^*] - p = \frac{\sum_{i=1}^{n} pf_i}{\sum_{i=1}^{n} w_i} - p = p\frac{\sum_{i=1}^{n} [f_i - w_i]}{\sum_{i=1}^{n} w_i}.$$

This estimator will be unbiased only if $\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} w_i$. In particular, p^* is unbiased if the correct weight function is used. The sample proportion (which occurs with the weight function $w(t) \equiv 1$), as expected, is usually biased, since $f_i \leq 1$.

Assuming the correct weight function has been chosen, the variance of p^* (Larsson & Sjögren, 1954) is

$$\text{Var}[p^*] = \frac{\sum_{i=1}^{n} pw_i(1 - pw_i)}{\left[\sum_{i=1}^{n} w_i\right]^2} = \frac{p}{\sum_{i=1}^{n} w_i}\left(1 - p\frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i}\right).$$

Some incorrect formulas for the variance which have been reported previously are

1) $\frac{p(1-p)}{n}$ (valid only when $w_i \equiv 1$) (Winokur et al, 1969).

2) $\frac{p(1-p)}{\sum_{i=1}^{n} w_i}$ which was pointed out previously as an incorrect formula by Larsson and Sjögren (1954). Risch (1983) also pointed out the Larsson and Sjögren formula is incorrect if one sets the estimate of p to 1 when $p^* > 1$.

With assumptions on the sequence w_i, such as they don't approach 0, p^* is asymptotically normal. Risch (1983) suggested using the asymptotic normality property for the construction of confidence intervals and hypothesis tests.

To avoid parameter estimates greater than one, Strömgren suggested the following modification to the estimator.

3.2.2 Modified Strömgren (1938):

$$p' = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (1 - w_i)x_i} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} (1 - x_i)w_i}.$$

This estimator always satisfies the condition of $p' \leq 1$, as the denominator is always greater that the number of affected people. However in situations where p^* is unbiased, p' will be have a negative bias since $p' \leq p^*$. This can is easily shown since

$$\sum_{i=1}^{n} w_i \leq \sum_{i=1}^{n} w_i + \sum_{i=1}^{n} (1 - w_i)x_i .$$

The difference between the Strömgren and the Modified Strömgren estimators appears to be an increasing function of p^*. This also suggests that when p^* is unbiased, the bias of p' is an increasing function of p.

A third proposal takes a maximum likelihood approach to the problem.

3.2.3 Maximum Likelihood (Risch, 1983):

The maximum likelihood estimate $\hat{p}$ is estimated by maximizing L(p) or log L(p), the likelihood or log-likelihood functions:

$$L(p) = \prod_{i=1}^{n} (pw_i)^{x_i}(1 - pw_i)^{1-x_i},$$

$$\log L(p) = \log p \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i \log w_i + \sum_{i=1}^{n} (1 - x_i) \log (1 - pw_i).$$

The estimate can be found by solving the following equation

$$\left.\frac{d \log L(p)}{dp}\right|_{\hat{p}} = \sum_{i=1}^{n} \left[\frac{x_i}{\hat{p}} - \frac{(1 - x_i)w_i}{1 - \hat{p}w_i}\right] = 0. \qquad (3)$$

This equation does not have a closed form solution but can be solved easily and quickly by the Newton-Raphson method. As the second derivative of log L(p) can be shown to be less than or equal to 0 for all $p \in (0,1]$, there is a unique $p \in [0,1]$ which maximizes log

L(p). As with the Strömgren estimator, using the solution to (3) can lead to an estimate greater than 1. This will occur when

$$\left.\frac{d \log L(p)}{dp}\right|_{p=1} = \sum_{i=1}^{n} \frac{x_i - w_i}{1 - w_i} > 0.$$

If this occurs, $\hat{p} = 1$.

Estimates of the variance of $\hat{p}$ can be obtained using the observed or the expected information. The observed information is:

$$I_O(p) = -\frac{d^2 \log L(p)}{dp^2} = \sum_{i=1}^{n} \left[\frac{x_i}{p^2} + \frac{(1-x_i)w_i^2}{(1-pw_i)^2}\right]$$

and the expected information is

$$I_E(p) = -E\left[\frac{d^2 \log L(p)}{dp^2}\right] = \sum_{i=1}^{n} \left[\frac{pw_i}{p^2} + \frac{(1-pw_i)w_i^2}{(1-pw_i)^2}\right] = \sum_{i=1}^{n} \left[\frac{w_i}{p} + \frac{w_i^2}{1-pw_i}\right].$$

When evaluated at the maximum likelihood estimate:

$$I_O(\hat{p}) = \sum_{i=1}^{n} \left[\frac{x_i}{\hat{p}^2} + \frac{(1-x_i)w_i^2}{(1-\hat{p}w_i)^2}\right]$$

$$= \sum_{i=1}^{n} \left[\frac{1}{\hat{p}}\frac{(1-x_i)w_i}{(1-\hat{p}w_i)} + \frac{(1-x_i)w_i^2}{(1-\hat{p}w_i)^2}\right]$$

$$= \sum_{i=1}^{n} \left[\left(\frac{1-x_i}{1-\hat{p}w_i}\right)\left(\frac{w_i}{\hat{p}} + \frac{w_i^2}{1-\hat{p}w_i}\right)\right]$$

$$I_E(\hat{p}) = \sum_{i=1}^{n} \left[\frac{w_i}{\hat{p}} + \frac{w_i^2}{1-\hat{p}w_i}\right].$$

It is not immediately obvious whether $I_E(\hat{p})^{-1}$ or $I_O(\hat{p})^{-1}$ would better estimate the variance; it appears that both would give similar values. The one advantage to using $I_O(\hat{p})^{-1}$ is that it is calculated when the Newton-Raphson method is used for estimation.

As the method of estimation is maximum likelihood, $\hat{p}$ is asymptotically normal allowing the construction of confidence intervals and hypothesis tests. Risch (1983) also proposes using the likelihood ratio test to compare estimates of p among two of more groups.

Misspecification of the weight function can lead to problems, as with p^*. As

$$E\left[\frac{d \log L(p)}{dp}\right] = \sum_{i=1}^{n}\left[f_i - \frac{(1-pf_i)w_i}{1-pw_i}\right] = \sum_{i=1}^{n}\left[\frac{f_i - w_i}{1-pw_i}\right]$$

may not be zero when an inappropriate weight function is chosen, $\hat{p}$ may not be a consistent estimator and one or both of the variance estimates may be poor.

Risch showed that $\hat{p}$ is more efficient than p^* and the efficiency of p^* relative to $\hat{p}$ is independent of the sample size. Risch also showed that:

$$\text{eff}(p^*) = \left\{1 + \frac{1}{W^2}\sum_{j=2}^{\infty}(W_{j+1}W - W_jW_2)\right\}^{-1}$$

where $W = \sum_{i=1}^{n} w_i$ and $W_j = \sum_{i=1}^{n} w_i^j$.

3.3 Results

Three different weight functions, as displayed in Table 3.1, were chosen to analyze the Alzheimer and the Winokur data sets. The plots of these functions are shown in Figures 3.1 and 3.2. The first two were chosen to roughly match the lower and upper observed ages of onset of the index cases and the relatives in the appropriate data sets.

Table 3.1 Weight Functions Used

	Alzheimer	Winokur
1) Half Risk	$w_1^A(t) = \begin{cases} 0 & t \le 34 \\ \frac{1}{2} & 35 \le t \le 90 \\ 1 & t \ge 91 \end{cases}$	$w_1^W(t) = \begin{cases} 0 & t \le 14 \\ \frac{1}{2} & 15 \le t \le 70 \\ 1 & t \ge 71 \end{cases}$
2) Uniform	$w_2^A(t) = \begin{cases} 0 & t \le 34 \\ \frac{t-34}{56} & 35 \le t \le 90 \\ 1 & t \ge 91 \end{cases}$	$w_2^W(t) = \begin{cases} 0 & t \le 14 \\ \frac{t-14}{56} & 15 \le t \le 70 \\ 1 & t \ge 71 \end{cases}$
3) Empiric CDF	$w_3^A(t)$ = empiric distribution function of age of onset of index cases, regardless of diagnostic criteria	$w_3^W(t)$ = empiric distribution function of age of onset of index cases

Figure 3.1: Alzheimer Weight Functions

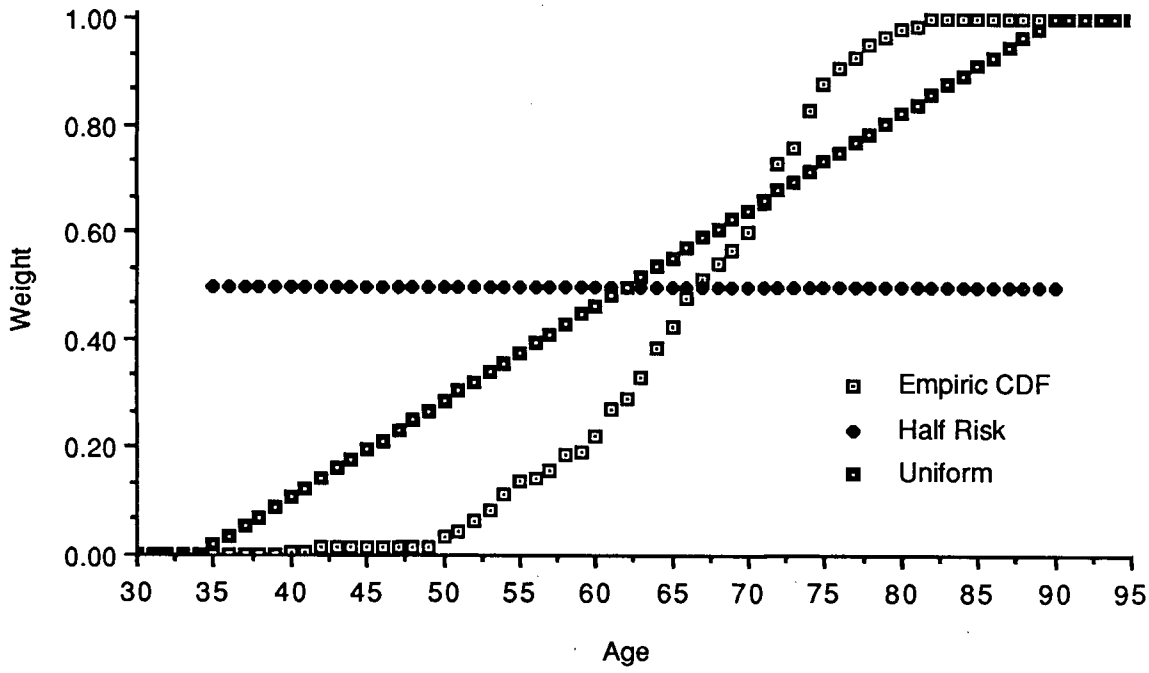
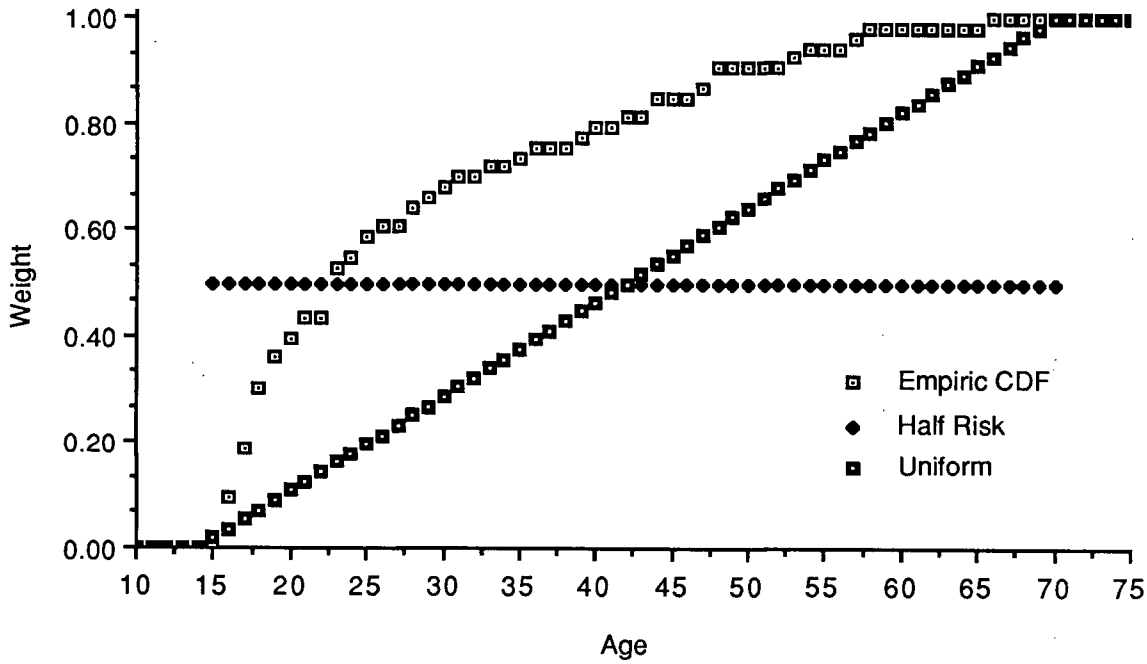


Figure 3.2: Winokur Weight Functions



As in Chapter 2, the Alzheimer data set was analyzed under the four criteria. One family (MM Pedigree, Figure 1.1) can greatly influence the results using the methods to be discussed in Chapter 4. It was felt that the data should be analyzed twice, with and then without this family, for the criteria containing the possible FAD families to see what effect this family has on this type of estimation. For the reanalysis, weight functions one and two were unaltered, but the empiric distribution function was modified to exclude the age of onset for the index case of the MM family.

Generally under each criterion for the Alzheimer data, the three weight functions seem to give similar risk estimates for each of the three estimators (Tables 3.1 - 3.4). The standard errors shown were calculated using the expected information. The increasing difference between the modified Strömgren and the Strömgren estimators with increasing p^* is suggested in the tables. The maximum likelihood and Strömgren estimators agree very closely under the Stringent without FAD, Stringent with FAD, and Relaxed Criteria for each of the weight functions. The largest differences between the maximum likelihood and the Strömgren estimators occurs under the FAD only criteria. However with the small number of relatives in the group, the differences are all less than one standard error. Also the exclusion of MM family appears to make little difference in the lifetime risk estimates with the biggest difference occurring under the FAD only criteria.

Except for the FAD only group, under these three weight functions, the lifetime risk for Alzheimer disease does not approach the 50% rate consistent with an autosomal dominant trait. In fact for these criteria, the lifetime risk estimates appear to be lower (though not significantly lower) than the age-specific risks to age 90 calculated by the product-limit estimators. This suggests that a poor set of weight functions may have been used.

For the Winokur data set, changing the weight function appears to make a great difference in the estimates (see Table 3.6). The estimates using the empirical distribution function are much lower than those for the other two weight functions. As the empirical distribution dominates the other weight functions for ages greater than 22, the large

differences are not surprising. It is not clear which of these three is the best choice for the weight function. However the ages of onset in the index cases appear to be less than the ages of onset in their relatives, suggesting that the empiric distribution may be a poor choice.

As it appears that a poor choice of onset distribution can lead to poor a estimate of lifetime risk, this suggests that a better procedure which will also estimate the age of onset distribution is needed. An extension to the maximum likelihood procedure of this chapter allowing the estimation of lifetime risk and the age of onset distribution will be discussed in the next chapter.

Table 3.2: Lifetime Risk Under Stringent without FAD Criteria

Weight Function	Method	Risk (SE)
Half Risk	Strömgren	0.093 (0.016)
	Modified Strömgren	0.081
	Maximum Likelihood	0.094 (0.016)
Uniform	Strömgren	0.077 (0.013)
	Modified Strömgren	0.067
	Maximum Likelihood	0.077 (0.013)
Empiric CDF	Strömgren	0.078 (0.013)
	Modified Strömgren	0.067
	Maximum Likelihood	0.078 (0.013)

Table 3.3: Lifetime Risk Under Stringent with FAD Criteria

Weight Function	Method	Risk (SE) With MM Family	Risk (SE) Without MM Family
Half Risk	Strömgren	0.131 (0.018)	0.122 (0.017)
	Modified Strömgren	0.109	0.102
	Maximum Likelihood	0.131 (0.018)	0.122 (0.017)
Uniform	Strömgren	0.109 (0.015)	0.100(0.014)
	Modified Strömgren	0.092	0.085
	Maximum Likelihood	0.109 (0.015)	0.101 (0.014)
Empiric CDF	Strömgren	0.111 (0.015)	0.103 (0.015)
	Modified Strömgren	0.093	0.086
	Maximum Likelihood	0.111 (0.015)	0.103 (0.015)

Table 3.4: Lifetime Risk Under Relaxed Criteria

Weight Function	Method	Risk (SE) With MM Family	Risk (SE) Without MM Family
Half Risk	Strömgren	0.189 (0.021)	0.180 (0.021)
	Modified Strömgren	0.146	0.140
	Maximum Likelihood	0.190 (0.021)	0.181 (0.021)
Uniform	Strömgren	0.157 (0.017)	0.149 (0.017)
	Modified Strömgren	0.123	0.117
	Maximum Likelihood	0.159 (0.018)	0.151 (0.017)
Empiric CDF	Strömgren	0.160 (0.018)	0.152 (0.017)
	Modified Strömgren	0.124	0.118
	Maximum Likelihood	0.160 (0.018)	0.153 (0.017)

Table 3.5: Lifetime Risk Under FAD Only Criteria

Weight Function	Method	Risk (SE) With MM Family	Risk (SE) Without MM Family
Half Risk	Strömgren	0.630 (0.124)	0.531 (0.124)
	Modified Strömgren	0.324	0.295
	Maximum Likelihood	0.597 (0.121)	0.509 (0.122)
Uniform	Strömgren	0.610 (0.114)	0.486 (0.110)
	Modified Strömgren	0.313	0.268
	Maximum Likelihood	0.554 (0.110)	0.476 (0.108)
Empiric CDF	Strömgren	0.677 (0.112)	0.528 (0.111)
	Modified Strömgren	0.332	0.279
	Maximum Likelihood	0.556 (0.108)	0.486 (0.108)

Table 3.6: Lifetime Risk for Winokur Data Set

Weight Function	Method	Risk (SE)
Half Risk	Strömgren	0.552 (0.076)
	Modified Strömgren	0.288
	Maximum Likelihood	0.511 (0.076)
Uniform	Strömgren	0.600 (0.083)
	Modified Strömgren	0.311
	Maximum Likelihood	0.575 (0.081)
Empiric CDF	Strömgren	0.355 (0.052)
	Modified Strömgren	0.215
	Maximum Likelihood	0.361 (0.052)

4 Lifetime Risk and Age of Onset Distribution Estimation

4.1 Background

The two methods discussed in the previous chapters, while computationally attractive, have major drawbacks. The product-limit procedures, which give good estimates of age-specific risks, cannot give the lifetime risk for disease unless possibly unreasonable assumptions are made. The fixed weight (age of onset) function approaches, though giving easily calculated lifetime risk estimates, can give poor estimates if an inappropriate weight function is used.

Risch (1983) suggested a maximum likelihood approach for calculating morbidity risks for diseases with late variable onset. It allows for the simultaneous estimation of lifetime risk and the age of onset distribution. This approach has also been used by Pericak-Vance et al (1983) to study the heterogeneity of age of onset of Huntington disease.

4.2 Model and Estimation

It is assumed that each relative belongs to one of two groups, susceptible or not susceptible with:

$$P[\text{susceptible}] = p = 1 - P[\text{not susceptible}].$$

For those in the susceptible group, it is assumed that their age of onset can be described by a distribution function $F(.|\underline{\theta})$ belonging to a class of distributions parametrized by $\underline{\theta} = (\theta_1,\ldots,\theta_k) \in \Theta$. Let the corresponding density function be $f(.|\underline{\theta})$.

For each relative i, let the random variable S_i denote the person's status and the random variable T_i denote the observation time.

$$\text{Let } s_i = \begin{cases} 0 & \text{if person i is unaffected} \\ 1 & \text{if person i is affected} \end{cases}$$

$$\text{Let } t_i = \begin{cases} \text{age at onset} & \text{if } s_i = 1 \text{ and age of onset is known} \\ \text{age at FH/Death} & \text{if } s_i = 1 \text{ and age of onset is unknown or if } s_i = 0 \end{cases}$$

where age at FH is the age of a live relative when the family history was collected .

Also, if appropriate, let $\underline{x}_i$ be a vector of covariates for person i. Examples of possible covariates are sex, information on other medical conditions, or age of onset of the index case. Then one or both of p and $\underline{\theta}$ could be functions of $\underline{x}$.

Then (with the possible dependence on $\underline{x}_i$ suppressed)

$$\begin{aligned} P[S_i = 1 \mid T_i = t_i] &= P[S_i = 1, \text{ susceptible} \mid T_i = t_i] + \\ &\qquad P[S_i = 1, \text{ not susceptible} \mid T_i = t_i] \\ &= P[\text{susceptible}]\, P[S_i = 1 \mid T_i = t_i, \text{ susceptible}] + \\ &\qquad P[\text{not susceptible}]\, P[S_i = 1 \mid T_i = t_i, \text{ not susceptible}] \\ &= pF(t_i|\underline{\theta}) \end{aligned}$$

and

$$\begin{aligned} P[S_i = 0 \mid T_i = t_i] &= P[S_i = 0, \text{ susceptible} \mid T_i = t_i] + \\ &\qquad P[S_i = 0, \text{ not susceptible} \mid T_i = t_i] \\ &= P[\text{susceptible}]\, P[S_i = 0 \mid T_i = t_i, \text{ susceptible}] + \\ &\qquad P[\text{not susceptible}]\, P[S_i = 0 \mid T_i = t_i, \text{ not susceptible}] \\ &= p(1 - F(t_i|\underline{\theta})) + (1-p) = 1 - pF(t_i|\underline{\theta}) \end{aligned}$$

Assume there are n relatives with relatives 1 to n_1 affected with known age of onset, $n_1 + 1$ to n_2 affected with unknown age of onset, and $n_2 + 1$ to n unaffected. Then the likelihood function and log likelihood functions are:

$$L(p,\underline{\theta}) = \prod_{i=1}^{n_1} pf(t_i|\underline{\theta}) \prod_{i=n_1+1}^{n_2} pF(t_i|\underline{\theta}) \prod_{i=n_2+1}^{n} [1 - pF(t_i|\underline{\theta})]$$

$$\log L(p,\underline{\theta}) = n_2 \log p + \sum_{i=1}^{n_1} \log f(t_i|\underline{\theta}) + \sum_{i=n_1+1}^{n_2} \log F(t_i|\underline{\theta}) + \sum_{i=n_2+1}^{n} \log[1 - pF(t_i|\underline{\theta})]$$

(It should be noted that there is a mistake in the first term of the log likelihood function (12) in Risch's paper. The correct term is $n_2 \log p$, not $(n_1 + n_2) \log p$). The maximum likelihood estimates for p, $\underline{\theta}$ (denoted by $\hat{p}$, $\hat{\underline{\theta}}$) can be calculated by standard procedures.

Now assume $f(t|\underline{\theta})$ has continuous first and second partial derivatives with respect to $\underline{\theta}$ and that the set $\{t \mid f(t|\underline{\theta}) > 0\}$ doesn't depend on $\underline{\theta}$. If the order of integration and differentiation can be changed, the first and second partial derivatives of $F(t|\underline{\theta})$ exist implying that $L(p,\underline{\theta})$ and $\log L(p,\underline{\theta})$ will also have continuous first and second partial derivatives. Then the maximum likelihood estimate $(\hat{p}, \hat{\underline{\theta}})$ satisfies the following system of equations (assuming it doesn't occur on the boundary of the parameter space):

$$\left.\frac{\partial \log L(p,\underline{\theta})}{\partial p}\right|_{\hat{p},\hat{\underline{\theta}}} = \frac{n_2}{\hat{p}} - \sum_{i=n_2+1}^{n}\left[\frac{F(t_i|\hat{\underline{\theta}})}{1-\hat{p}F(t_i|\hat{\underline{\theta}})}\right] = 0$$

$$\left.\frac{\partial \log L(p,\underline{\theta})}{\partial \theta_i}\right|_{\hat{p},\hat{\underline{\theta}}} = \sum_{i=1}^{n_1}\frac{\partial \log f(t_i|\hat{\underline{\theta}})}{\partial \theta_i} + \sum_{i=n_1+1}^{n_2}\frac{\partial \log F(t_i|\hat{\underline{\theta}})}{\partial \theta_i} - \sum_{i=n_2+1}^{n}\left[\frac{\hat{p}}{1-\hat{p}F(t_i|\hat{\underline{\theta}})}\frac{\partial F(t_i|\hat{\underline{\theta}})}{\partial \theta_i}\right] = 0 \; ; \; i = 1,\ldots,k$$

For most if not all choices of F, this system of equations will not have a closed form solution and must be solved numerically. The Newton-Raphson method appears useful here as it has good convergence properties and gives an estimate of the variance-covariance matrix of the parameter estimates. However for some families of distributions, such as the gamma, some of the partial derivatives are difficult to calculate, suggesting a quasi-Newton-Raphson approach would be more appropriate.

One problem with a Newton-Raphson type approach is that the procedure may not converge if poor initial estimates are chosen. Some of the time this can be overcome some of the time by scaling the difference between iterates or by replacing intermediate values outside the parameter space with values contained in the parameter space. Otherwise, the likelihood function will have to be investigated to find better initial estimates. It appears that the quality of the choice of initial estimates is less important for some choices of the age of onset distribution such as the logistic.

In some cases, it is possible that the maximum likelihood estimate may occur on the boundary of the parameter space, for example $\hat{p} = 1$. This situation can be suggested by the

Newton-Raphson iterations and must be confirmed by examining the likelihood function. In many cases when this happens, the parameter estimates can be found easily by fixing the appropriate parameters to their respective boundary values and estimating the rest by Newton-Raphson.

Let I be the observed information matrix. When the estimates are in the interior of the parameter space, I^{-1} provides an estimate of the variance-covariance of the parameter estimates. When the estimates of one or more parameters occurs on the boundary, the inverse of the observed information matrix may not be an appropriate estimate of the variance-covariance matrix as the required relationships

$$\left.\frac{\partial \log L(p,\underline{\theta})}{\partial p}\right|_{\hat{p},\hat{\underline{\theta}}} = 0\,;$$

$$\left.\frac{\partial \log L(p,\underline{\theta})}{\partial \theta_i}\right|_{\hat{p},\hat{\underline{\theta}}} = 0\,;\ i = 1,\ldots,k$$

will not necessarily hold.

When the variance-covariance is well-defined, approximate confidence sets for the parameters can be constructed using standard multivariate normal theory, since maximum likelihood estimates are asymptotically normal. In particular, an approximate $100(1-\alpha)\%$ confidence interval for p is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{[I^{-1}]_{0,0}}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile for a standard normal distribution and $[I^{-1}]_{0,0}$ is the entry in position (0,0) of I^{-1}.

Once the parameter estimates have been calculated, age-specific risks like those calculated using the methods of Chapter 2 can also be estimated.

Let r(t) = P[onset by age t]

= P[susceptible] P[onset by age t | susceptible]

= p $F(t|\underline{\theta})$.

Then the maximum likelihood estimate of r(t) is $\widehat{r(t)} = \hat{p}F(t|\hat{\underline{\theta}})$ and the following estimate of $\mathrm{Var}[\widehat{r(t)}]$ can be obtained by the delta method:

$$\left[F(t|\hat{\underline{\theta}}),\ \hat{p}\frac{\partial F(t|\hat{\underline{\theta}})}{\partial \theta_1},\ \ldots,\ \hat{p}\frac{\partial F(t|\hat{\underline{\theta}})}{\partial \theta_k} \right] I^{-1} \begin{bmatrix} F(t|\hat{\underline{\theta}}) \\ \hat{p}\frac{\partial F(t|\hat{\underline{\theta}})}{\partial \theta_1} \\ \vdots \\ \hat{p}\frac{\partial F(t|\hat{\underline{\theta}})}{\partial \theta_k} \end{bmatrix}$$

Breitner et al. (1988) used a multi-stage procedure which approximates the general maximum likelihood procedure. They started by estimating age-specific risks R_t by the Kaplan-Meier procedure as discussed in Chapter 2. Then they assumed that the age of onset distribution was gamma (discussed in Section 4.3) and found the least-squares estimates of the lifetime risk and the parameters of the gamma distribution (denoted a and m) by finding the values of p,a,m which minimize:

$$\sum_{t \in T} \left(R_t - pF(t|a,m)\right)^2$$

where T is the set of ages where at least one onset occurred and R_t is the age-specific risk at age t. Finally they proceeded to estimate the parameters of the gamma distribution by maximum likelihood by using the estimate of lifetime risk obtained from the least-squares procedure as if it were the known value of p. They claimed, due to technical reasons which were not stated, maximum likelihood fails to estimate lifetime risk, implying another procedure such as theirs is needed.

4.3 Age of Onset Distributions

Five different classes of distribution were chosen to model age of onset for the two data sets.

4.3.1. Multiple Hit (Gamma):

$$f(t|a,m) = \frac{a^m t^{m-1}}{\Gamma(m)} e^{-at}: \qquad t\geq 0,\ a>0,\ m = 1,2,3,\ldots$$

$$E[T] = \frac{m}{a}; \qquad Var[T] = \frac{m}{a^2}$$

Under this model the time to onset is the time for m (possible hypothetical) hits or shocks of frequency a to occur. It is assumed that the times between hits are independent exponential random variables with mean = a^{-1}. This model has been hypothesized by Breitner et al. (1986) as a possible model for describing the age of onset in Alzheimer disease.

The estimation procedure used under the Multiple Hit model is slightly different than in the general situation presented earlier since the scale parameter is constrained to take on only integer values for interpretation purposes.

Let $L_*(m) = \sup_{\{p,a\}} L(p,a,m)$ and p(m),a(m) be defined such that $L(p(m),a(m),m) = L_*(m)$. The values p(m) and a(m) can be calculated by Newton-Raphson for each m. Then define m^* to be the smallest m to satisfy $L_*(m^*) \geq L_*(m)$ for all $m \neq m^*$. Then the constrained maximum likelihood estimates for p, a, m are $\hat{p} = p(m^*)$, $\hat{a} = a(m^*)$, and $\hat{m} = m^*$.

By constraining m to be an integer, the gradient of log L(p,a,m) at the maximum likelihood estimate is not the zero vector. Because of this, the inverse of the observed information matrix may not be an appropriate estimate of the variance-covariance matrix, and therefore no estimate of the variance-covariance matrix will be reported.

4.3.2. Incubation (Lognormal):

$$f(t|\mu,\sigma) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right): \qquad t\geq 0,\ -\infty<\mu<\infty,\ \sigma>0$$

$$E[T] = \exp(\mu+0.5\sigma^2); \qquad Var[T] = \exp(2\mu+\sigma^2)(\exp(\sigma^2) - 1)$$

Sartwell (1950) showed that the incubation times for many infectious diseases could be described by a lognormal distribution. Later it was shown (Armenian and Khoury, 1981) that the age of onset for some non-infectious conditions, such as Huntington disease, could

also be modeled using a lognormal distribution. By examining four data sets from the literature, Horner (1987) suggested that Alzheimer disease may also satisfy the lognormal model.

4.3.3. Logistic:

$$f(t|\alpha,\beta) = \beta \frac{e^{\alpha+\beta t}}{(1+e^{\alpha+\beta t})^2}: \qquad t\geq 0,\ -\infty<\alpha<\infty,\ \beta>0$$

$$E[T] = -\frac{\alpha}{\beta}; \qquad Var[T] = \frac{\pi^2}{3\beta^2}$$

The logistic model was chosen since it has the computational advantage of having closed form expressions for $f(t|\alpha,\beta)$, $F(t|\alpha,\beta)$ and their first and second partial derivatives. Also, this distribution has been used in modelling the age of onset in hereditary polyposis coli (Morales et al, 1984)

4.3.4. Normal:

$$f(t|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right): \qquad t\geq 0,\ -\infty<\mu<\infty,\ \sigma>0$$

$$E[T] = \mu; \qquad Var[T] = \sigma^2$$

This class of distributions was chosen for two reasons. The first is that it was used by Risch (1983) in modelling the Winokur data set and and by Pericak-Vance et al. (1983) to model age of onset in Huntington's disease. Also the normal distribution is a limiting case for both of the gamma and lognormal distributions.

4.3.5. Normal with Covariate:

$$f(t|a,b,\sigma,x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{t-(a+bx)}{\sigma}\right)^2\right): \qquad t\geq 0,\ -\infty<a<\infty,\ -\infty<b<\infty,\ \sigma>0$$

$$E[T|x] = a + bx; \qquad Var[T|x] = \sigma^2$$

where x_i = age of onset in index case of person i

In some Alzheimer families it appears that the ages of onset are very similar (Sadovnick et al, 1988). One such example is shown in the pedigree of family MM (see Figure 1.1). It seems that a better fit may occur by trying to incorporate a distinct age of

onset distribution for each family. This type of model is desirable from a genetic counselling point of view as the smaller variation in the ages of onset, which would result in this model, would lead to better statements about risk. Assume the mean age of onset for person i can be modelled as $\mu_i = a + bx_i$. A model which would be easy to interpret would have $a \approx 0$ and $b \approx 1$. This would imply that the index case's age of onset is a good predictor for the age of onset in other family members. If the parameter estimates were different than this, in particular $b \neq 1$, interpretation seems to be more difficult.

As the method of estimation is maximum likelihood, the improvement of this model over the normal model can be assessed by looking at the likelihood ratio statistic testing the hypothesis of whether b is zero. The age of onset was not clear in a few of the index cases. So the two models could be compared, these few families were deleted from the data and the estimation for both models was done on the reduced data set.

4.4 Results

Under each of the four diagnostic criteria for the Alzheimer data set all five models were fit. The parameter estimates with the estimates of the variance-covariance matrices where appropriate are shown in Tables 4.1 - 4.4 and plots of the age-specific risk are shown in Figures 4.1 to 4.7. The maximum likelihood estimates of the means and standard deviations shown in the tables under each age of onset model were calculated by:

$$\widehat{\text{mean}} = E[T|\hat{\underline{\theta}}]; \quad \widehat{SD} = \sqrt{Var[T|\hat{\underline{\theta}}]}\ .$$

Also shown in the tables are the values of the log-likelihood under each of the models. These can be used as rough guides to suggest distributions which may not be appropriate, but as the models are not nested, conventional hypothesis tests cannot be constructed.

For the Alzheimer data set, the age of onset distribution for one family (MM pedigree; see Figure 1.1) is clearly different than that of the rest of the study group. When this family is included in the data set the parameter estimates appear to be greatly influenced by this family. In fact in some cases the estimate of lifetime risk is one, which does not appear to make biological sense. The inclusion of this very young onset family also leads to the

estimate of the mean age of onset to be much larger than generally recognized, which seem to be counterintuitive. Therefore, where appropriate, the data was analyzed twice for each of the criteria for which the MM family would be included, once with the family included, then with the family excluded. The estimates calculated with the MM family excluded appear to be in much greater agreement with the published literature.

Except for the FAD Only criteria analysis, the following relationship of the estimates holds:

$\hat{p}$ (logistic) < $\hat{p}$ (normal) < $\hat{p}$ (gamma) < $\hat{p}$ (lognormal)

$\widehat{mean}$ (logistic) < $\widehat{mean}$ (normal) < $\widehat{mean}$ (gamma) < $\widehat{mean}$ (lognormal)

$\widehat{SD}$ (logistic) < $\widehat{SD}$ (normal) < $\widehat{SD}$ (gamma) < $\widehat{SD}$ (lognormal)

Even though the four basic age of onset distributions can give estimates for lifetime risk which appear to be quite different, the age specific risks up to age 90, the upper bound on the data, appear to be almost the same. This can be seen in Figures 4.1 to 4.7. These age-specific risk estimates are also very similar to the product-limit estimates discussed in Chapter 2. One example showing the close agreement is given in Figure 4.8.

Although the parameters calculated under a given criteria when the MM family is included and excluded can give very different values, the estimates of the age specific risk are again very similar. Two examples showing this are displayed in Figures 4.9 and 4.10.

Using the age of onset as a predictor always appears to be a better model than the smaller normal model. The presence of the MM family with its very strong age correlation does not affect the decision about which is a more appropriate model, though it does influence the lifetime risk estimate. The effect of this family is much smaller here than for the other choices of distribution.

As for the product-limit and fixed age of onset methods, there is no indication of a 50% risk by age 90 as has been suggested in the other studies mentioned, except under the FAD only criteria. As these are families which are believed to represent the genetic form of the disease, this is to be expected. If the lognormal distribution is appropriate, it appears that under the relaxed criteria the data is consistent with a lifetime risk of 50%, as 0.5 is contained

in the approximate 95% confidence interval for p. However this is the only case where this occurs.

The parameter estimates for the Winokur data set are shown in Table 4.5 and the age-specific risks are shown in Figure 4.11. For this data set, the estimates for the lifetime risk and the mean and standard deviation for the age of onset distribution seem to be the same for the four basic models. However for this data set, using the age of onset of the index cases as a predictor doesn't give a significantly better fit. This suggests that the correlation of the ages of onset within families is not very strong in this data set.

Table 4.1: Parameter Estimates Under Stringent without FAD Criteria

Model	
Gamma	$\hat{p} = 0.186$; $\hat{a} = 0.501$; $\hat{m} = 41$ $\widehat{mean} = 81.791$; $\widehat{SD} = 12.774$ log L = -205.830
Lognormal	$\hat{p} = 0.210$; $\hat{\mu} = 4.424$; $\hat{\sigma} = 0.175$ $\widehat{mean} = 84.692$; $\widehat{SD} = 14.969$ Var = $\begin{bmatrix} 0.013 & 0.012 & 0.004 \\ 0.012 & 0.012 & 0.005 \\ 0.004 & 0.005 & 0.002 \end{bmatrix}$ log L = -206.053
Logistic	$\hat{p} = 0.146$; $\hat{\alpha} = -14.548$; $\hat{\beta} = 0.189$ $\widehat{mean} = 77.121$; $\widehat{SD} = 9.615$ Var = $\begin{bmatrix} 0.002 & 0.026 & -0.001 \\ 0.026 & 5.663 & -0.081 \\ -0.001 & -0.081 & 0.001 \end{bmatrix}$ log L = -205.004
Normal	$\hat{p} = 0.157$ $\hat{\mu} = 78.436$; $\hat{\sigma} = 10.346$ $\widehat{mean} = 78.436$; $\widehat{SD} = 10.346$ Var = $\begin{bmatrix} 0.003 & 0.178 & 0.070 \\ 0.178 & 17.333 & 7.259 \\ 0.070 & 7.259 & 4.444 \end{bmatrix}$ log L = -205.490
Normal with Covariate	$\hat{p} = 0.152$; $\hat{a} = 46.885$; $\hat{b} = 0.453$; $\hat{\sigma} = 9.289$ Var = $\begin{bmatrix} 0.002 & 0.020 & 0.002 & 0.048 \\ 0.020 & 149.158 & -2.178 & 1.143 \\ 0.002 & -2.178 & 0.034 & 0.0582 \\ 0.048 & 1.143 & 0.058 & 3.362 \end{bmatrix}$ log L = -196.406 -2 log Λ = 5.791

Table 4.2: Parameter Estimates Under Stringent with FAD Criteria

Model	With MM Family	Without MM Family
Gamma	$\hat{p} = 1$; $\hat{a} = 0.076$; $\hat{m} = 10$ $\widehat{mean} = 130.849$; $\widehat{SD} = 41.378$ log L = -318.001	$\hat{p} = 0.208$; $\hat{a} = 0.502$; $\hat{m} = 40$ $\widehat{mean} = 79.673$; $\widehat{SD} = 12.597$ log L = -278.671
Lognormal	$\hat{p} = 1$; $\hat{\mu} = 4.874$; $\hat{\sigma} = 0.391$ $\widehat{mean} = 141.205$; $\widehat{SD} = 57.447$ log L = -318.489	$\hat{p} = 0.227$; $\hat{\mu} = 4.390$; $\hat{\sigma} = 0.174$ $\widehat{mean} = 81.841$; $\widehat{SD} = 14.367$ $\text{Var} = \begin{bmatrix} 0.008 & 0.007 & 0.003 \\ 0.007 & 0.007 & 0.003 \\ 0.003 & 0.003 & 0.001 \end{bmatrix}$ log L = -278.921
Logistic	$\hat{p} = 0.221$;$\hat{\alpha} = -9.794$; $\hat{\beta} = 0.123$ $\widehat{mean} = 79.552$; $\widehat{SD} = 14.739$ $\text{Var} = \begin{bmatrix} 0.004 & 0.029 & -0.001 \\ 0.029 & 1.641 & -0.025 \\ -0.001 & -0.025 & 0.0004 \end{bmatrix}$ log L = -315.489	$\hat{p} = 0.176$; $\hat{\alpha} = -13.512$; $\hat{\beta} = 0.177$ $\widehat{mean} = 76.174$; $\widehat{SD} = 10.225$ $\text{Var} = \begin{bmatrix} 0.001 & 0.021 & -0.0004 \\ 0.021 & 3.586 & -0.052 \\ -0.0004 & -0.052 & 0.001 \end{bmatrix}$ log L = -278.121
Normal	$\hat{p} = 0.359$; $\hat{\mu} = 90.109$; $\hat{\sigma} = 19.235$ $\widehat{mean} = 90.109$; $\widehat{SD} = 19.235$ $\text{Var} = \begin{bmatrix} 0.061 & 3.357 & 1.094 \\ 3.357 & 195.598 & 66.069 \\ 1.094 & 66.069 & 24.787 \end{bmatrix}$ log L = -316.840	$\hat{p} = 0.183$; $\hat{\mu} = 76.891$; $\hat{\sigma} = 10.468$ $\widehat{mean} = 76.891$; $\widehat{SD} = 10.468$ $\text{Var} = \begin{bmatrix} 0.002 & 0.127 & 0.052 \\ 0.127 & 11.315 & 4.938 \\ 0.052 & 4.938 & 3.283 \end{bmatrix}$ log L = -278.376
Normal with Covariate	$\hat{p} = 0.207$; $\hat{a} = 14.976$; $\hat{b} = 0.930$; $\hat{\sigma} = 10.626$ $\text{Var} = \begin{bmatrix} 0.002 & 0.020 & 0.001 & 0.047 \\ 0.020 & 77.138 & -1.139 & 0.456 \\ 0.001 & -1.139 & 0.019 & 0.057 \\ 0.047 & 0.456 & 0.057 & 2.873 \end{bmatrix}$ log L = -293.114 -2 log Λ = 34.613	$\hat{p} = 0.179$; $\hat{a} = 41.264$; $\hat{b} = 0.516$; $\hat{\sigma} = 9.138$ $\text{Var} = \begin{bmatrix} 0.002 & -0.001 & 0.001 & 0.033 \\ -0.001 & 106.491 & -1.581 & 0.121 \\ 0.001 & -1.581 & 0.025 & 0.043 \\ 0.032 & 0.121 & 0.043 & 2.217 \end{bmatrix}$ log L = -266.893 -2 log Λ = 10.131

Table 4.3: Parameter Estimates Under Relaxed Criteria

Model	With MM Family	Without MM Family
Gamma	$\hat{p} = 1$; $\hat{a} = 0.115$; $\hat{m} = 13$ $\widehat{mean} = 113.536$; $\widehat{SD} = 31.489$ log L = -387.195	$\hat{p} = 0.376$; $\hat{a} = 0.424$; $\hat{m} = 36$ $\widehat{mean} = 84.905$; $\widehat{SD} = 14.151$ log L = -347.229
Lognormal	$\hat{p} = 1$; $\hat{\mu} = 4.742$; $\hat{\sigma} = 0.328$ $\widehat{mean} = 121.069$; $\widehat{SD} = 40.833$ log L = -388.362	$\hat{p} = 0.438$; $\hat{\mu} = 4.468$; $\hat{\sigma} = 0.190$ $\widehat{mean} = 88.752$; $\widehat{SD} = 16.974$ $\text{Var} = \begin{bmatrix} 0.046 & 0.022 & 0.007 \\ 0.022 & 0.011 & 0.004 \\ 0.007 & 0.004 & 0.002 \end{bmatrix}$ log L = -347.502
Logistic	$\hat{p} = 0.351$;$\hat{\alpha} = -10.586$; $\hat{\beta} = 0.128$ $\widehat{mean} = 82.513$; $\widehat{SD} = 14.137$ $\text{Var} = \begin{bmatrix} 0.009 & 0.041 & -0.001 \\ 0.041 & 1.344 & -0.020 \\ -0.001 & -0.020 & 0.0003 \end{bmatrix}$ log L = -383.843	$\hat{p} = 0.281$; $\hat{\alpha} = -13.568$; $\hat{\beta} = 0.171$ $\widehat{mean} = 79.095$; $\widehat{SD} = 10.574$ $\text{Var} = \begin{bmatrix} 0.003 & 0.032 & -0.001 \\ 0.032 & 2.543 & -0.037 \\ -0.001 & -0.037 & 0.001 \end{bmatrix}$ log L = -346.465
Normal	$\hat{p} = 0.726$; $\hat{\mu} = 97.372$; $\hat{\sigma} = 19.845$ $\widehat{mean} = 97.372$; $\widehat{SD} = 19.845$ $\text{Var} = \begin{bmatrix} 0.343 & 9.027 & 2.642 \\ 9.027 & 245.855 & 74.220 \\ 2.642 & 74.220 & 24.068 \end{bmatrix}$ log L = -385.745	$\hat{p} = 0.310$; $\hat{\mu} = 80.765$; $\hat{\sigma} = 11.300$ $\widehat{mean} = 80.765$; $\widehat{SD} = 11.300$ $\text{Var} = \begin{bmatrix} 0.007 & 0.288 & 0.112 \\ 0.288 & 14.676 & 6.062 \\ 0.112 & 6.062 & 3.321 \end{bmatrix}$ log L = -346.879
Normal with Covariate	$\hat{p} = 0.337$; $\hat{a} = 25.391$; $\hat{b} = 0.804$; $\hat{\sigma} = 10.977$ $\text{Var} = \begin{bmatrix} 0.004 & 0.162 & 0.000 & 0.065 \\ 0.162 & 69.487 & -0.929 & 3.427 \\ 0.000 & -0.929 & 0.014 & 0.002 \\ 0.065 & 3.427 & 0.002 & 2.275 \end{bmatrix}$ log L = -372.418 -2 log Λ = 31.789	$\hat{p} = 0.320$; $\hat{a} = 48.241$; $\hat{b} = 0.454$; $\hat{\sigma} = 9.805$ $\text{Var} = \begin{bmatrix} 0.004 & 0.133 & 0.0002 & 0.059 \\ 0.133 & 84.865 & -1.167 & 3.029 \\ 0.0002 & -1.167 & 0.018 & 0.0002 \\ 0.059 & 3.029 & 0.0002 & 1.964 \end{bmatrix}$ log L = -344.751 -2 log Λ = 10.567

Table 4.4: Parameter Estimates under FAD Only Criteria

Model	With MM Family	Without MM Family
Gamma	$\hat{p} = 1$; $\hat{a} = 0.090$; $\hat{m} = 8$ $\widehat{mean} = 89.027$; $\widehat{SD} = 31.476$ log L = -86.032	$\hat{p} = 0.554$; $\hat{a} = 0.766$; $\hat{m} = 54$ $\widehat{mean} = 70.457$; $\widehat{SD} = 9.588$ log L = -59.129
Lognormal	$\hat{p} = 1$; $\hat{\mu} = 4.428$; $\hat{\sigma} = 0.402$ $\widehat{mean} = 90.817$; $\widehat{SD} = 38.075$ log L = -85.995	$\hat{p} = 0.562$; $\hat{\mu} = 4.251$; $\hat{\sigma} = 0.141$ $\widehat{mean} = 70.841$; $\widehat{SD} = 10.035$ $Var = \begin{bmatrix} 0.020 & 0.005 & 0.003 \\ 0.005 & 0.003 & 0.002 \\ 0.003 & 0.002 & 0.002 \end{bmatrix}$ log L = -59.168
Logistic	$\hat{p} = 0.686$;$\hat{\alpha} = -6.700$; $\hat{\beta} = 0.094$ $\widehat{mean} = 71.034$; $\widehat{SD} = 19.229$ $Var = \begin{bmatrix} 0.036 & 0.093 & -0.003 \\ 0.093 & 2.202 & -0.036 \\ -0.003 & -0.036 & 0.001 \end{bmatrix}$ log L= -85.084	$\hat{p} = 0.552$; $\hat{\alpha} = -12.764$; $\hat{\beta} = 0.181$ $\widehat{mean} = 70.387$; $\widehat{SD} = 10.002$ $Var = \begin{bmatrix} 0.016 & 0.100 & -0.002 \\ 0.100 & 11.248 & -0.168 \\ -0.002 & -0.168 & 0.003 \end{bmatrix}$ log L = -59.350
Normal	$\hat{p} = 0.698$; $\hat{\mu} = 71.376$; $\hat{\sigma} = 18.892$ $\widehat{mean} = 71.376$; $\widehat{SD} = 18.892$ $Var = \begin{bmatrix} 0.058 & 2.000 & 0.982 \\ 2.000 & 92.988 & 44.219 \\ 0.982 & 44.219 & 31.871 \end{bmatrix}$ log L = -85.923	$\hat{p} = 0.542$; $\hat{\mu} = 69.945$; $\hat{\sigma} = 9.055$ $\widehat{mean} = 69.945$; $\widehat{SD} = 9.055$ $Var = \begin{bmatrix} 0.016 & 0.180 & 0.093 \\ 0.180 & 9.895 & 3.801 \\ 0.093 & 3.801 & 5.623 \end{bmatrix}$ log L = -59.138
Normal with Covariate	$\hat{p} = 0.566$; $\hat{a} = 4.084$; $\hat{b} = 1.000$; $\hat{\sigma} = 8.245$ $Var = \begin{bmatrix} 0.014 & 0.014 & 0.002 & 0.069 \\ 0.014 & 89.037 & 1.439 & 0.239 \\ 0.002 & -1.439 & 0.025 & 0.033 \\ 0.069 & 0.239 & 0.033 & 3.219 \end{bmatrix}$ log L = -76.416 -2 log Λ = 19.115	$\hat{p} = 0.539$; $\hat{a} = 39.477$; $\hat{b} = 0.464$; $\hat{\sigma} = 7.623$ $Var = \begin{bmatrix} 0.015 & 0.044 & 0.002 & 0.058 \\ 0.043 & 245.385 & -3.752 & 2.0612 \\ 0.001 & -3.752 & 0.059 & 0.002 \\ 0.058 & 2.062 & 0.002 & 3.575 \end{bmatrix}$ log L = -57.631 -2 log Λ = 3.015

Table 4.5: Parameter Estimates For Winokur Data Set

Model	
Gamma	$\hat{p} = 0.507$; $\hat{a} = 0.175$; $\hat{m} = 7$ $\widehat{mean} = 40.0$; $\widehat{SD} = 15.1$ log L = -156.791
Lognormal	$\hat{p} = 0.539$; $\hat{\mu} = 3.666$; $\hat{\sigma} = 0.429$ $\widehat{mean} = 42.873$; $\widehat{SD} = 19.294$ $\text{Var} = \begin{bmatrix} 0.019 & 0.019 & 0.009 \\ 0.019 & 0.027 & 0.012 \\ 0.009 & 0.012 & 0.008 \end{bmatrix}$ log L = -156.301
Logistic	$\hat{p} = 0.457$; $\hat{\alpha} = -5.264$; $\hat{\beta} = 0.145$ $\widehat{mean} = 36.346$; $\widehat{SD} = 12.524$ $\text{Var} = \begin{bmatrix} 0.006 & 0.007 & -0.001 \\ 0.007 & 0.662 & -0.019 \\ -0.001 & -0.019 & 0.001 \end{bmatrix}$ log L = -159.477
Normal	$\hat{p} = 0.478$ $\hat{\mu} = 37.801$; $\hat{\sigma} = 12.468$ $\widehat{mean} = 37.801$; $\widehat{SD} = 12.468$ $\text{Var} = \begin{bmatrix} 0.007 & 0.154 & 0.068 \\ 0.154 & 10.275 & 4.142 \\ 0.068 & 4.142 & 4.297 \end{bmatrix}$ log L = -158.843
Normal with Covariate	$\hat{p} = 0.437$; $\hat{a} = 27.248$; $\hat{b} = 0.278$,; $\hat{\sigma} = 10.815$ $\text{Var} = \begin{bmatrix} 0.006 & 0.210 & -0.003 & 0.058 \\ 0.210 & 37.099 & -0.826 & 5.583 \\ -0.003 & -0.826 & 0.023 & -0.073 \\ 0.058 & 5.583 & -0.073 & 3.481 \end{bmatrix}$ log L = -157.449 $-2 \log \Lambda = 2.787$

Figure 4.1: Probability of Being Affected Under Stringent without FAD Criteria

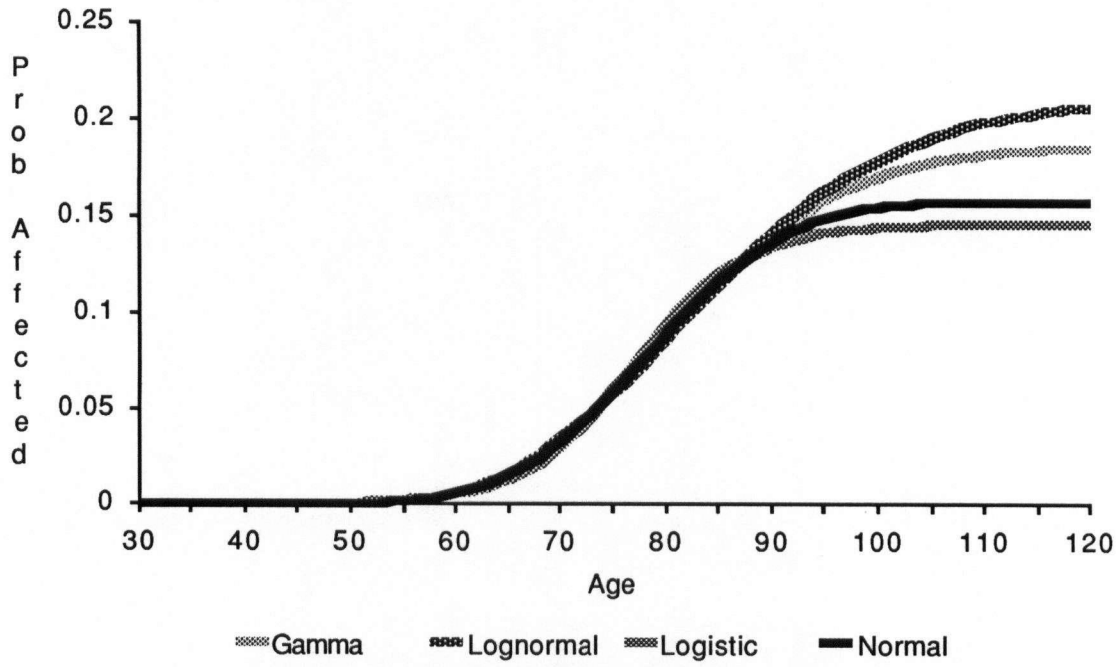


Figure 4.2: Probability of Being Affected Under Stringent with FAD Criteria (MM Family Included)

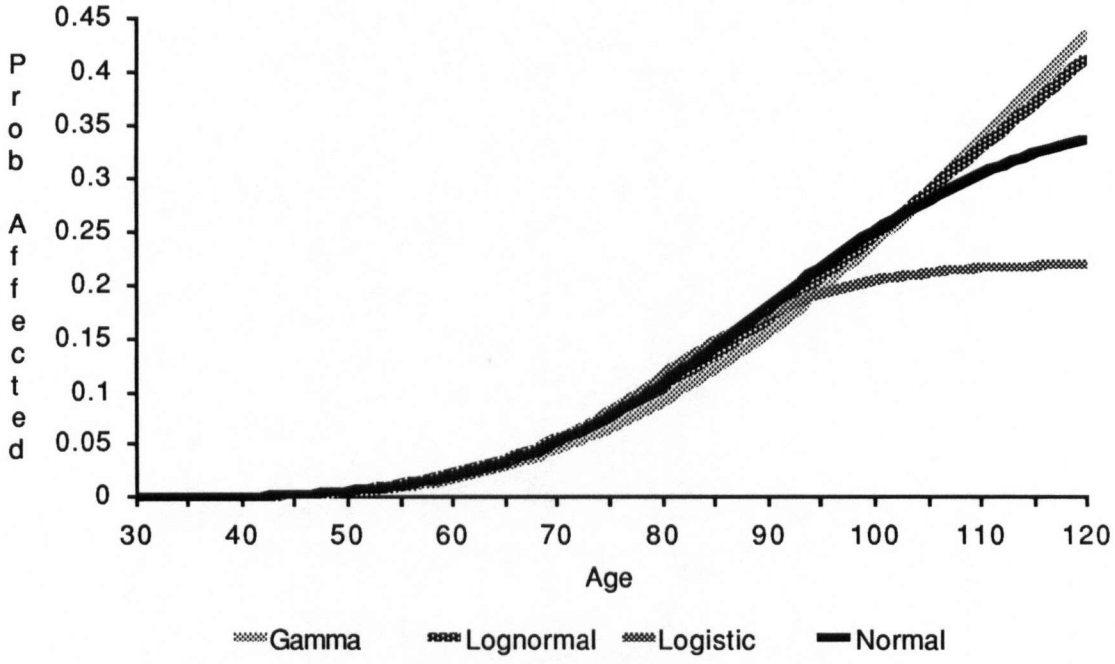


Figure 4.3: Probability of Being Affected Under Stringent with FAD Criteria (MM Family Excluded)

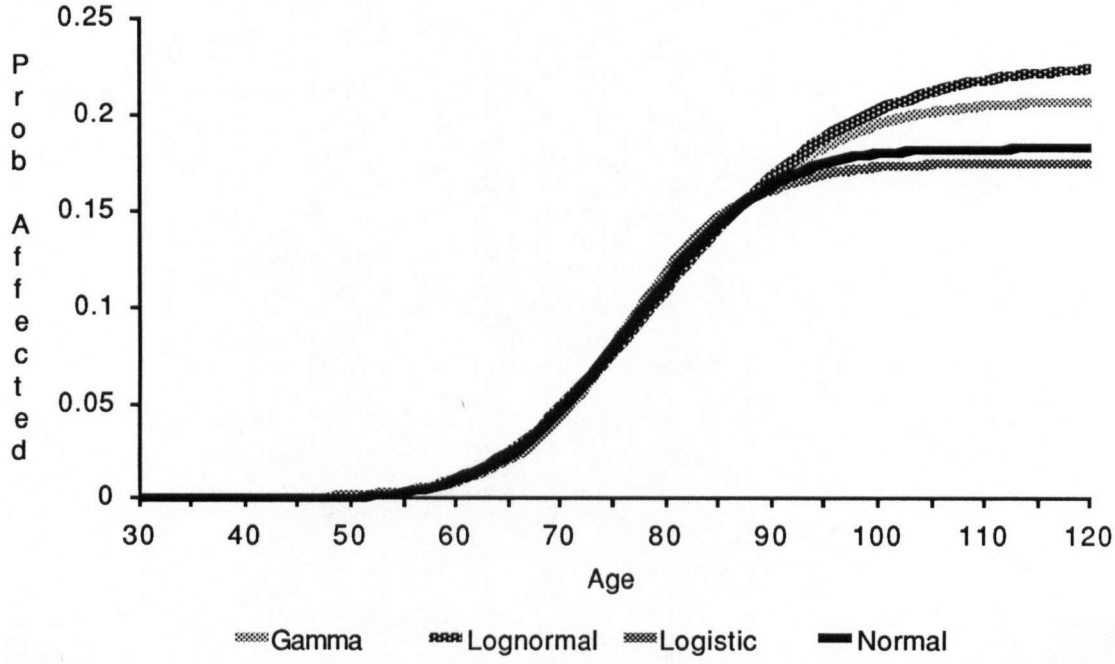


Figure 4.4: Probability of Being Affected Under Relaxed Criteria
(MM Family Included)

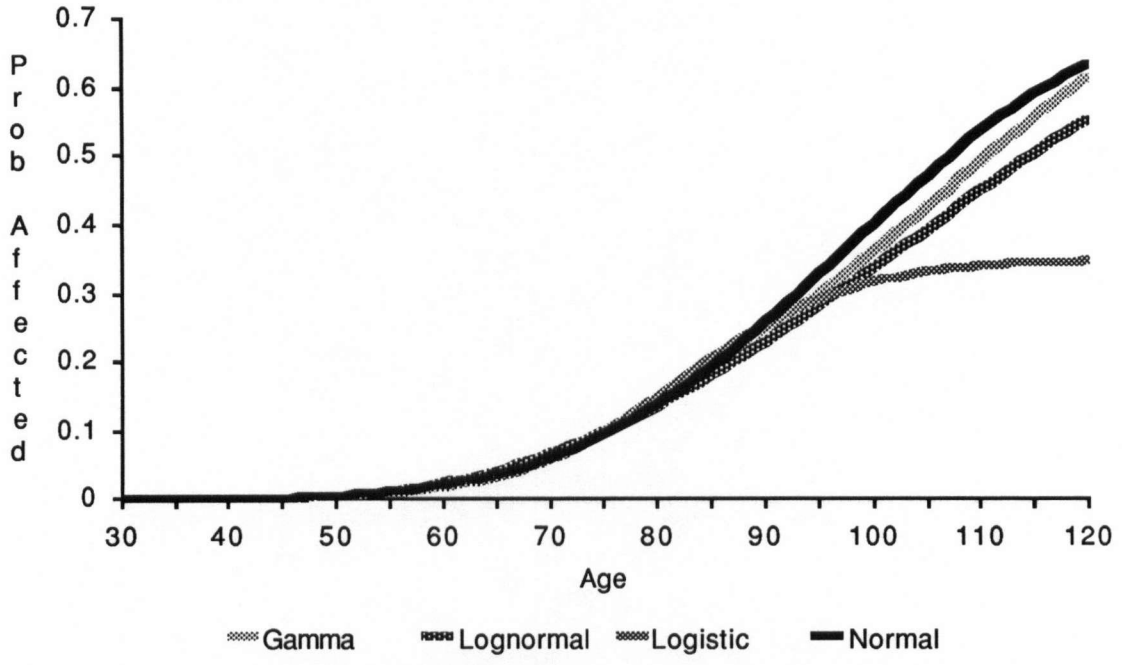


Figure 4.5: Probability of Being Affected Under Relaxed Criteria
(MM Family Excluded)

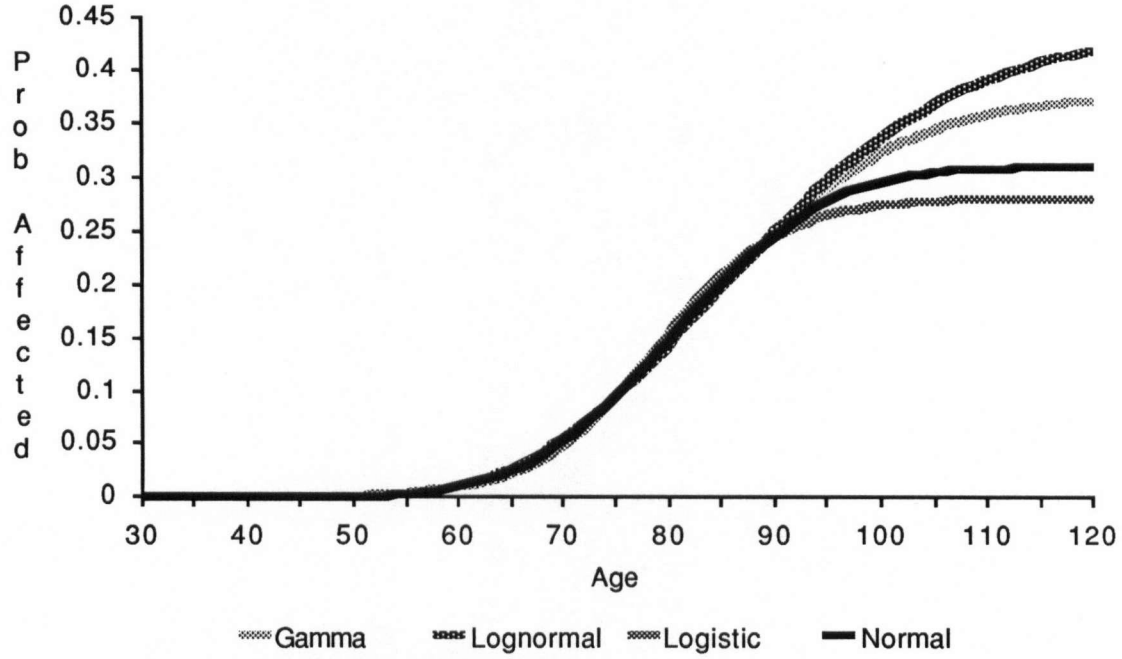


Figure 4.6: Probability of Being Affected Under FAD Only Criteria
(MM Family Included)

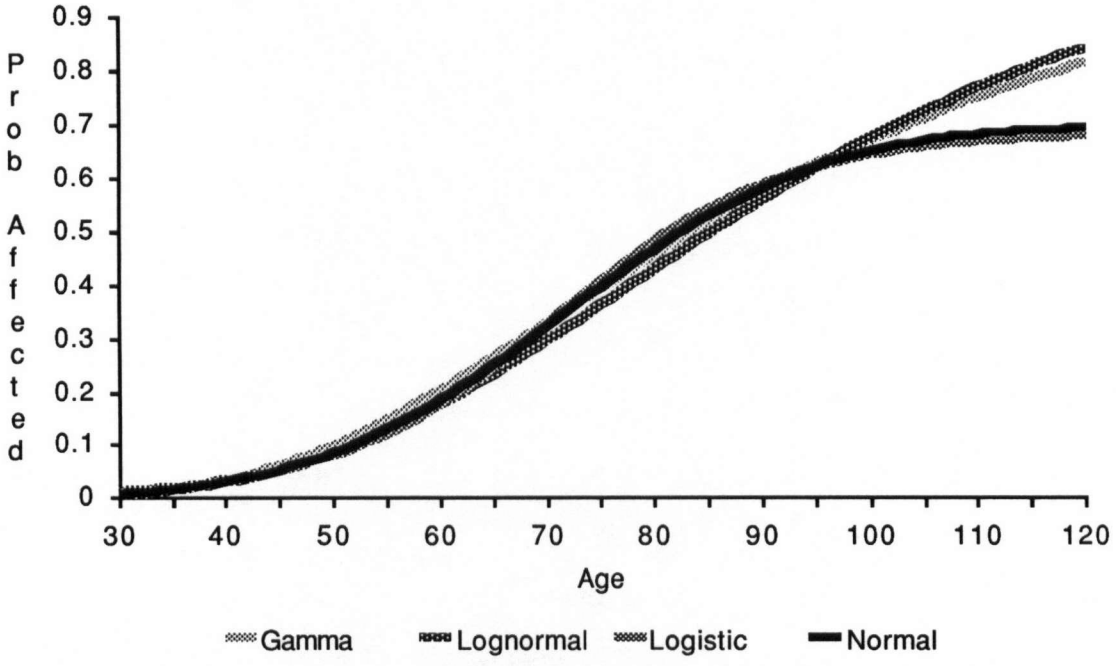


Figure 4.7: Probability of Being Affected Under FAD Only Criteria
(MM Family Excluded)

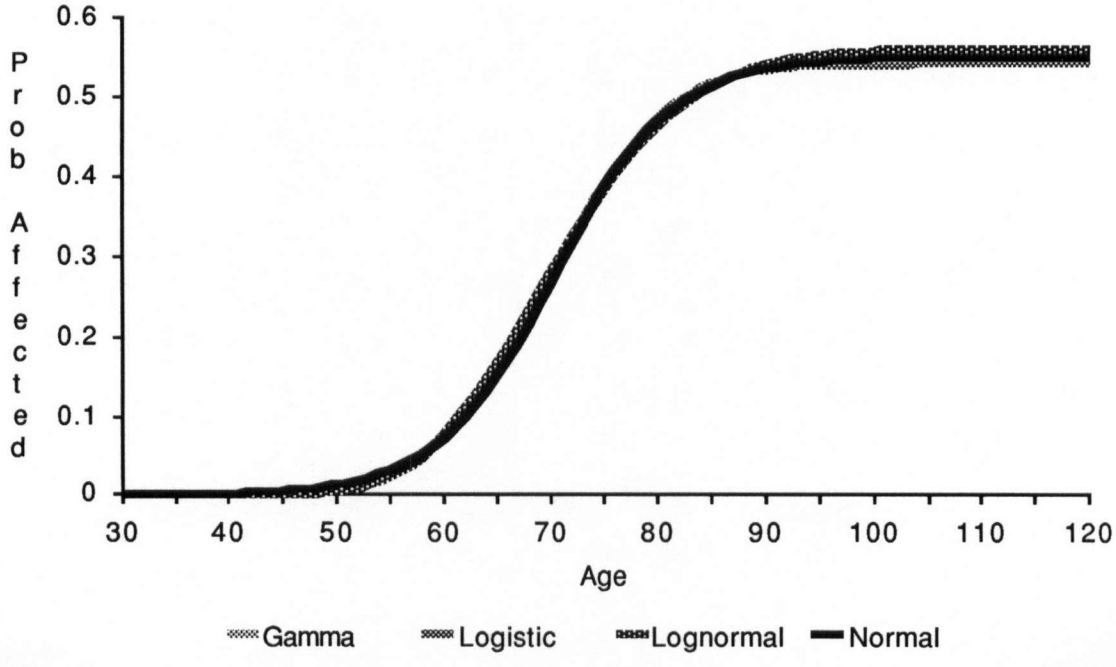


Figure 4.8: Probability of Being Affected Under Stringent without FAD Criteria with Life-Table Estimate

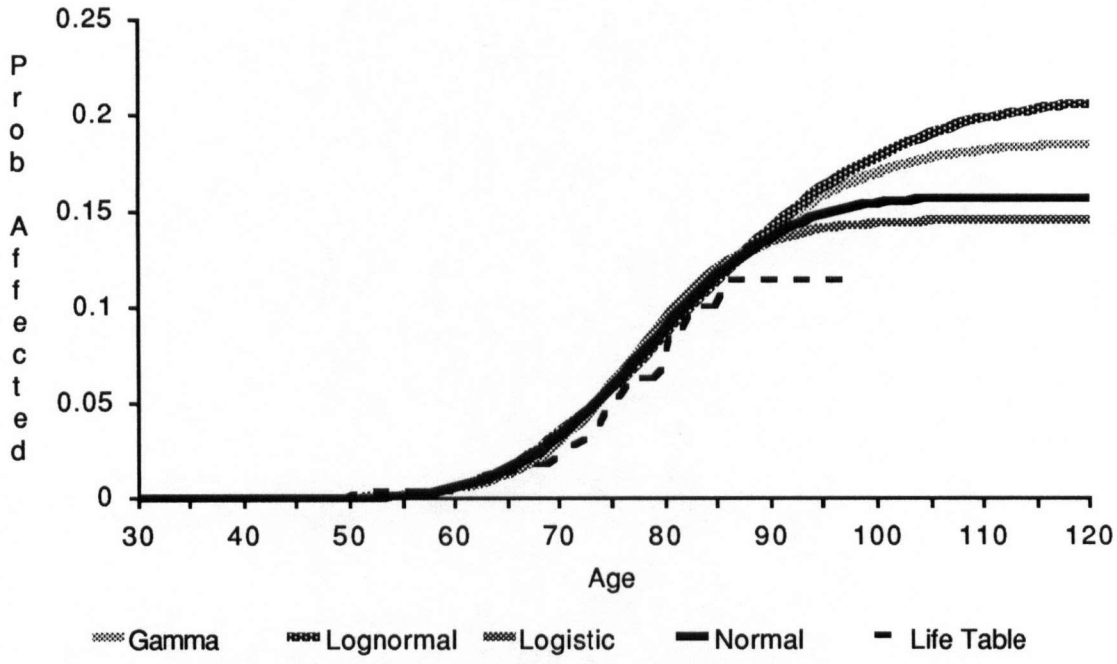


Figure 4.9: Probability of Being Affected Under Relaxed Criteria (Effect of MM Family with Normal Age of Onset)

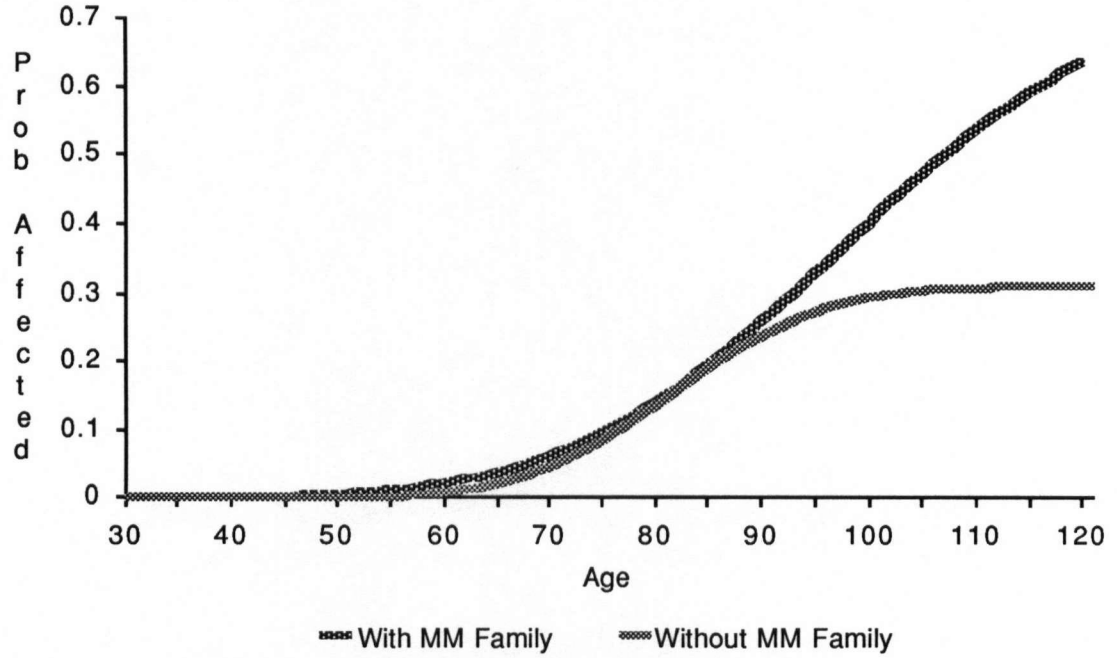


Figure 4.10: Probability of Being Affected Under Stringent with FAD Criteria
(Effect of MM Family with Gamma Age of Onset)

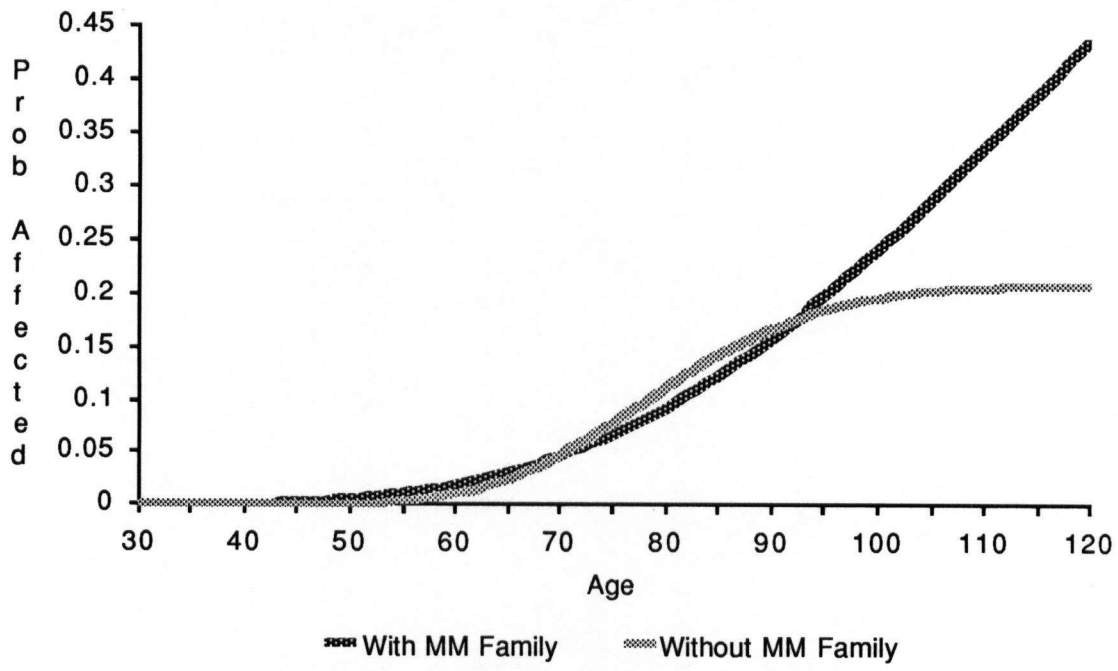
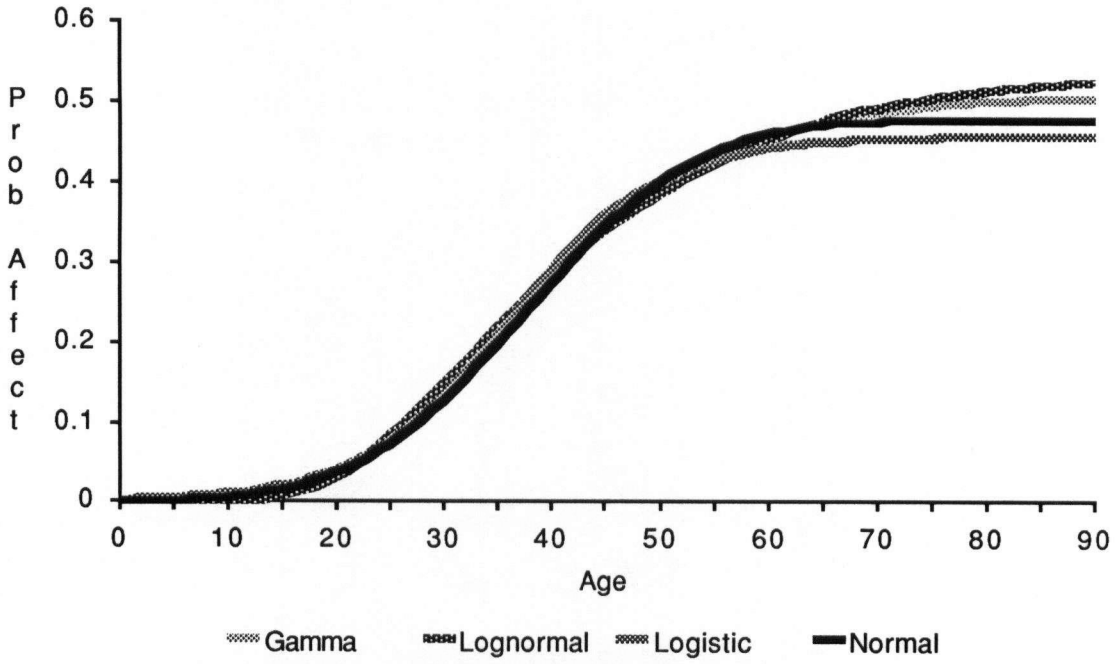


Figure 4.11: Probability of Being Affected For Winokur Data Set



5 SIMULATION STUDY

5.1 Background

As the method of estimation discussed in Chapter 4 is maximum likelihood, it is known that the estimators are consistent and asymptotically normal when the correct class of distributions is chosen. But due to the complex structure of this mixture model, making statements about the rates of convergence of the estimators is difficult. Also, it is not clear what will happen when an incorrect class of distributions is chosen to fit the data. Monte-Carlo simulations can be used to address these questions.

5.2 Simulation Conditions

There were four factors which were varied in this simulation study: the proportion of people susceptible, the size of the simulated data sets, the class of distribution functions for the age of onset, the mean and variances of the age of onset and censoring distribution. It was felt that these factors could have the greatest affect on the estimates. The choices for the factors are shown in Table 5.1.

The one factor which was not varied was the use of a gamma distribution for the censoring process. It was felt that the mean and standard deviation, not the class of the censoring distribution, would be the more important factor. The mean, and to a lesser extent the standard deviation, of the censoring distribution should be chosen to be similar to the mean and standard deviation of the age of onset distribution. This property has been observed in the two data sets considered here and also appears to occur in Multiple Sclerosis. In particular, the choices should not imply most people will be censored before reaching their age of onset, which doesn't seem to occur in practice. This undesirable condition can happen when the mean of the censoring distribution is set too low.

Table 5.1: Parameters of the Simulation Study

p	Data Set Size
0.5	150, 500
0.25	150, 500
0.15	500

Age of Onset		Censoring	
Mean	Standard Deviation	Mean	Standard Deviation
80	10	65	20
55	10	65	20
35	10	35	10

Class of the Age of Onset Distributions: Logistic, Normal, Gamma, Lognormal.

Class of the Censoring Distribution: Gamma

Data Sets per configuration: 200.

The random numbers required for the onset and censoring processes were generated using the parametrizations described in Chapter 4, where the parameters of each distribution were set to match the desired mean and standard deviations. However, in the gamma case, to match the assumption used in Chapter 4 that the shape parameter is an integer the following parametrization was used:

$$m = \text{int}\left[\left(\frac{\text{mean}}{\text{standard deviation}}\right)^2\right], \quad a = \frac{\text{mean}}{(\text{standard deviation})^2},$$

where int[r] is the integer part of a real number r. Using this parametrization implies that the mean and standard deviation actually used for the gamma distribution are slightly smaller than the desired values when the nominal values for the mean are 55 and 35. The values for

m in these two cases are 30 and 12 instead of 30.25 and 12.25. The true values for the mean and standard deviation under the gamma case are:

Nominal Mean	True Mean	True Standard Deviation
80	80	10
55	54.5455	9.9586
35	34.2857	9.8974

These small differences in the means and standard deviations shouldn't affect the conclusions.

The possible configuration with p = 0.15 and 150 people in the data set was not considered in the simulations as some preliminary runs suggested that this was an unstable situation which would not be particularly informative. The problem appeared to be that not enough affected people were contained in these simulated data sets leading to divergence even when the correct distribution was being used for estimation.

Each simulated data set was examined under the four age of onset distributions used for generating the data. Two forms of the gamma distribution were used in estimation; with and without the assumption that the shape parameter m is an integer.

5.3 Results

Three parameters of the processes have been estimated for each configuration: the lifetime risk, and the mean and standard deviation of the age of onset. It was decided to investigate the mean and standard deviation as these can be compared across various distributions, whereas the natural parameters of the age of onset distributions cannot. In a few cases, the Newton-Raphson procedure did not converge. Sample averages (with standard errors) using the cases that converged for each of these three parameters under the 60 configurations are shown in Tables 5.2 to 5.13. The averages which differ from the true value by more than two standard errors, indicating possible bias, are in bold type. As a large number of comparisons are done (180 just to test for unbiasedness of the three parameters

under the correct model), some of these deviations could be due to random variation. Assuming that all the estimators were unbiased, using this two standard error rule would lead to about 5% of the cases appearing to be biased.

The restricted and unrestricted gamma models appear to give almost the same estimates under all of the configurations considered. The major difference appears to be in the estimate of the shape parameter, with the restricted estimate set to the nearest integer of the unrestricted estimate in almost all cases. Because of this, only the properties under the unrestricted model will be discussed.

The first parameter considered is the lifetime risk for disease. Irrespective of what the true class of the distribution is, it appears that the expected value of the lifetime risk will satisfy the relation

$$E[\hat{p} \mid \text{logistic fit}] \leq E[\hat{p} \mid \text{normal fit}] \leq E[\hat{p} \mid \text{gamma fit}] \leq E[\hat{p} \mid \text{lognormal fit}].$$

This holds for all but six of the sixty configurations. In four cases the average under the logistic is larger than the average under the normal, and in two cases the average under the gamma is larger than the average under the lognormal. However, the differences are very small and could to be due to random variation.

In general, it appears that when the correct class of age of onset distribution is chosen for estimation, the estimates appear to be approximately unbiased as only four out of the sixty (6.67%) averages differ significantly from the true value of p. In all four cases the difference appears to be small. The apparent unbiasedness is to be expected from the consistency of maximum likelihood estimates.

However, when the incorrect distribution is chosen, biases in the estimates appear. The normal and logistic distributions appear to do fairly well when trying to approximate the other. This is not very surprising as the shapes of these distributions are quite similar; in particular both are symmetric about their means. Also both of these distributions do fairly well when trying to estimate in the gamma and lognormal settings. It appears that the maximum size of the bias is about 10%. The use of the gamma or lognormal distributions when the true distribution is either logistic or normal leads to positive and occasionally large

biases. In fact, when the gamma and lognormal models are used to estimate a logistic or normal situation with a mean age of onset of 35, the Newton-Raphson procedure won't converge and suggests that the maximum likelihood estimate in these cases is 1. For these cases, the true expected value of the estimator is probably higher than is shown in the tables. When the gamma and lognormal are used to estimate the other, the biases again appear to be small with the maximum size of the bias also about 10%.

The largest biases seem to occur when a mean of 35 is chosen. This should not be surprising since the distributions are most dissimilar in this case. As the mean increases, the gamma and lognormal distributions become more symmetric and look more like the normal and logistic distributions.

The expected values for the estimates of the mean age of onset appear to have a similar property to the expected values for the estimates of the lifetime risk. Regardless of the true distribution, it appears:

$$E[\widehat{\text{mean}}|\text{logistic fit}] \leq E[\widehat{\text{mean}}|\text{normal fit}] \leq E[\widehat{\text{mean}}|\text{gamma fit}]$$
$$\leq E[\widehat{\text{mean}}|\text{lognormal fit}].$$

This relationship holds for all but four configurations, with the average under the logistic larger than the average under the normal in three cases, and the average under the gamma larger than the average under the lognormal in one case. Again the differences are small and could be due to random variation.

When the correct class of distributions is used, the estimate of the mean also appears to be approximately unbiased, with the averages from only two of the sixty configurations differing from the true value by more than two standard errors. As the size of the apparent bias appears to be small, this could be due to random variation.

As with the estimates of lifetime risk, choosing the incorrect distribution appears to lead to a biased estimate of the mean age of onset. The logistic and normal both appear to do fairly well in all cases, in particular when trying to estimate the other. Again the maximum bias appears to be less than 10%, with the size decreasing with increasing mean. The gamma and lognormal distribution do fairly well except for trying to estimate the logistic and

normal distribution when the mean age of onset is 35. This again is due, at least in part, to the convergence problems.

For lifetime risk and mean age of onset estimates, there is a mild suggestion that bias is more likely when the number of relatives is 500 rather that 150. This could be due to the fact that the greater amount of data leading to much less variation in the estimates. As there should be approximately three to four times more affected people in the data sets with 500 relatives, less variation should be expected.

The structure of the average values observed for the lifetime risk and the mean age of onset breaks down when the standard deviation is considered. Similar to before, independent of the true class of onset distributions:

$$E[\widehat{SD}|\text{logistic fit}] \leq E[\widehat{SD}|\text{gamma fit}] \leq E[\widehat{SD}|\text{lognormal fit}] \text{ and}$$

$$E[\widehat{SD}|\text{normal fit}] \leq E[\widehat{SD}|\text{gamma fit}] \leq E[\widehat{SD}|\text{lognormal fit}].$$

However $E[\widehat{SD}|\text{logistic fit}] \leq E[\widehat{SD}|\text{normal fit}]$ only seems to hold when the true age of onset distribution is logistic with a mean of 80 or 35. The opposite holds when the true distribution is normal, gamma, or lognormal, or when the true distribution is logistic with a mean of 55. As the size of the differences in the average estimate between the normal and logistic distributional assumptions when the true age of onset distribution is logistic with a mean of 55 are small, the reversal in ordering could be due to chance, but the consistent pattern seems to suggest otherwise.

When the correct distribution is chosen, the estimate often appears to have a negative bias (18/60 = 30%). There was also one situation with an apparent positive bias. It appears bias is least likely when the true mean is 55 and most likely when the true mean is 80. Since

$$P[\text{affected}|\text{mean} = 80] \leq P[\text{affected}|\text{mean}=35] \leq P[\text{affected}|\text{mean} = 55]$$

appears to hold, this suggests that the number affected in the data set determines the size of the bias. This may also explain why the lognormal (2/15) and gamma (3/15) distributions appear to have fewer bias problems than the logistic (5/15) and the normal (8/15). In this case the interaction between the age of onset and censoring distributions also appears to be involved, as it determines the number of affected people.

When the wrong distribution is chosen, a biased estimate appears in most situations. The maximum bias appears to be approximately 30%, which occurs when the mean age of onset is 35. As with the estimates for lifetime risk and mean age of onset, the values in the tables for the gamma and lognormal distributions are probably underestimated when the true age of onset distribution is logistic or normal with a mean of 35.

For the classes of distributions chosen, an important factor in determining the bias of lifetime risk, mean and standard deviation of the age of onset distribution and the probability of the Newton-Raphson procedure converging appears to be the relative size of the left tails of the true and fitting distribution. These simulations suggest that the heavier the tail of the fitting distributions, the lower the estimate of lifetime risk and mean age of onset. However if the tail is too light, the lifetime risk estimate will be larger. Also the estimated age of onset distribution will shift to the right suggesting larger ages of onset. The importance of the size of the left tail agrees with the effect the MM family had on the estimates in Chapter 4. This family gives the true age of onset distribution for the sample a heavier left tail, leading to the expected shifts in the estimates under this hypothesis.

These simulations suggest that choosing a distribution with a heavier left tail, such as the logistic, can lead to robustness against outliers with the possible cost of slightly underestimating the lifetime risk and the average age of onset. This finding is in disagreement with Risch's (1983) statement based on the Winokur data set that the estimate of the lifetime risk didn't particularly depend on the choice of age of onset distribution. However, Risch did make the very important point that one should try to characterize the choice of onset distribution as carefully as possible prior to analysis.

Table 5.2: Average of Estimates for p (Generated by Logistic)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	0.4945 (0.0081)	0.4992 (0.0084)	0.5060 (0.0087)	**0.5193 (0.0091)**
0.5	80	500	0.4981 (0.0038)	0.5058 (0.0039)	**0.5252 (0.0046)**	**0.5391 (0.0050)**
0.25	80	150	0.2581 (0.0060)	0.2590 (0.0062)	**0.2731 (0.0071)**	**0.2746 (0.0074)**
0.25	80	500	0.2525 (0.0037)	0.2561 (0.0039)	**0.2656 (0.0042)**	**0.2739 (0.0047)**
0.15	80	500	0.1479 (0.0027)	0.1511 (0.0028)	**0.1596 (0.0039)**	**0.1659 (0.0049)**
0.5	55	150	0.4997 (0.0043)	0.5003 (0.0043)	**0.5140 (0.0048)**	**0.5252 (0.0052)**
0.5	55	500	0.4995 (0.0021)	0.5005 (0.0022)	**0.5129 (0.0026)**	**0.5236 (0.0030)**
0.25	55	150	0.2465 (0.0036)	0.2472 (0.0036)	0.2529 (0.0041)	0.2601 (0.0054)
0.25	55	500	0.2520 (0.0018)	0.2525 (0.0018)	**0.2578 (0.0020)**	**0.2654 (0.0023)**
0.15	55	500	0.1501 (0.0015)	0.1502 (0.0015)	**0.1536 (0.0017)**	**0.1581 (0.0018)**
0.5	35	150	0.5047 (0.0066)	**0.5206 (0.0072)**	**0.5933 (0.0112)**	**0.6474 (0.0156)**
0.5	35	500	0.4987 (0.0033)	**0.5159 (0.0036)**	**0.6759 (0.0117)**	**0.7696 (0.0141)**
0.25	35	150	0.2580 (0.0059)	**0.2666 (0.0064)**	**0.3354 (0.0127)**	**0.3508 (0.0131)**
0.25	35	500	0.2504 (0.0029)	**0.2596 (0.0032)**	**0.3620 (0.0114)**	**0.4809 (0.0182)**
0.15	35	500	0.1517 (0.0020)	**0.1575 (0.0023)**	**0.2308 (0.0103)**	**0.2946 (0.1573)**

Table 5.3: Average of Estimates for p (Generated by Normal)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	0.5105 (0.0087)	0.5168 (0.0092)	**0.5222 (0.0094)**	**0.5312 (0.0096)**
0.5	80	500	0.4929 (0.0045)	0.4988 (0.0046)	0.5093 (0.0049)	**0.5160 (0.0050)**
0.25	80	150	0.2557 (0.0068)	0.2570 (0.0069)	**0.2666 (0.0074)**	**0.2690 (0.0081)**
0.25	80	500	0.2492 (0.0035)	0.2518 (0.0035)	0.2552 (0.0037)	**0.2607 (0.0038)**
0.15	80	500	0.1518 (0.0028)	0.1526 (0.0028)	0.1552 (0.0030)	**0.1582 (0.0031)**
0.5	55	150	**0.5150 (0.0041)**	**0.5145 (0.0041)**	**0.5234 (0.0044)**	**0.5326 (0.0050)**
0.5	55	500	0.4998 (0.0025)	0.4997 (0.0025)	**0.5066 (0.0026)**	**0.5139 (0.0026)**
0.25	55	150	0.2504 (0.0036)	0.2500 (0.0035)	0.2544 (0.0037)	**0.2586 (0.0038)**
0.25	55	500	0.2532 (0.0018)	0.2531 (0.0018)	**0.2567 (0.0019)**	**0.2605 (0.0019)**
0.15	55	500	**0.1536 (0.0015)**	**0.1537 (0.0015)**	**0.1561 (0.0016)**	**0.1583 (0.0016)**
0.5	35	150	0.4909 (0.0068)	0.5006 (0.0071)	**0.5622 (0.0109)**	**0.6355 (0.0141)**
0.5	35	500	**0.4902 (0.0031)**	0.5019 (0.0034)	**0.6124 (0.0077)**	**0.7429 (0.0114)**
0.25	35	150	0.2466 (0.0048)	0.2543 (0.0056)	**0.3168 (0.0117)**	**0.3521 (0.0145)**
0.25	35	500	0.2494 (0.0029)	0.2557 (0.0032)	**0.3272 (0.0077)**	**0.4166 (0.0118)**
0.15	35	500	**0.1439 (0.0021)**	0.1469 (0.0022)	**0.1953 (0.0065)**	**0.2499 (0.0109)**

Table 5.4: Average of Estimates for p (Generated by Gamma)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	0.4858 (0.0077)	0.4929 (0.0079)	0.4995 (0.0082)	0.5006 (0.0085)
0.5	80	500	**0.4870 (0.0045)**	0.4941 (0.0050)	0.5011 (0.0047)	0.5064 (0.0049)
0.25	80	150	0.2498 (0.0065)	0.2515 (0.0066)	0.2596 (0.0067)	0.2569 (0.0070)
0.25	80	500	**0.2416 (0.0035)**	0.2451 (0.0036)	0.2493 (0.0037)	0.2505 (0.0038)
0.15	80	500	0.1527 (0.0028)	0.1544 (0.0028)	**0.1569 (0.0029)**	**0.1581 (0.0030)**
0.5	55	150	0.4970 (0.0039)	0.4986 (0.0039)	0.5031 (0.0040)	0.5072 (0.0040)
0.5	55	500	**0.4919 (0.0022)**	**0.4941 (0.0022)**	0.4980 (0.0022)	0.5017 (0.0023)
0.25	55	150	0.2503 (0.0033)	0.2513 (0.0034)	0.2531 (0.0034)	0.2551 (0.0034)
0.25	55	500	0.2467 (0.0018)	0.2480 (0.0018)	0.2497 (0.0019)	0.2516 (0.0019)
0.15	55	500	0.1486 (0.0015)	0.1495 (0.0015)	0.1505 (0.0015)	0.1515 (0.0016)
0.5	35	150	**0.4699 (0.0050)**	**0.4766 (0.0051)**	0.5085 (0.0065)	**0.5487 (0.0085)**
0.5	35	500	**0.4614 (0.0030)**	**0.4692 (0.0031)**	0.4971 (0.0036)	**0.5280 (0.0045)**
0.25	35	150	**0.2309 (0.0041)**	**0.2336 (0.0042)**	0.2527 (0.0063)	**0.2684 (0.0068)**
0.25	35	500	**0.2330 (0.0022)**	**0.2377 (0.0023)**	0.2528 (0.0028)	**0.2717 (0.0035)**
0.15	35	500	**0.1405 (0.0020)**	**0.1437 (0.0021)**	0.1544 (0.0028)	**0.1709 (0.0047)**

Table 5.5: Average of Estimates for p (Generated by Lognormal)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	0.4921 (0.0084)	0.4941 (0.0080)	0.5013 (0.0084)	0.5059 (0.0086)
0.5	80	500	**0.4763 (0.0040)**	**0.4844 (0.0041)**	**0.4891 (0.0041)**	0.4930 (0.0042)
0.25	80	150	0.2421 (0.0070)	0.2445 (0.0071)	0.2567 (0.0074)	0.2493 (0.0075)
0.25	80	500	0.2469 (0.0033)	0.2505 (0.0034)	0.2530 (0.0035)	0.2550 (0.0035)
0.15	80	500	0.1483 (0.0026)	0.1513 (0.0027)	0.1519 (0.0028)	0.1540 (0.0029)
0.5	55	150	**0.4844 (0.0039)**	**0.4875 (0.0039)**	**0.4912 (0.0040)**	0.4929 (0.0040)
0.5	55	500	**0.4915 (0.0023)**	**0.4950 (0.0023)**	0.4980 (0.0023)	0.5005 (0.0023)
0.25	55	150	0.2493 (0.0037)	0.2505 (0.0037)	0.2532 (0.0037)	0.2538 (0.0038)
0.25	55	500	**0.2451 (0.0017)**	0.2474 (0.0017)	0.2480 (0.0018)	0.2499 (0.0017)
0.15	55	500	0.1500 (0.0014)	0.1512 (0.0014)	0.1520 (0.0014)	0.1527 (0.0015)
0.5	35	150	**0.4596 (0.0059)**	**0.4697 (0.0062)**	0.4929 (0.0071)	0.5124 (0.0082)
0.5	35	500	**0.4566 (0.0027)**	**0.4674 (0.0029)**	**0.4855 (0.0032)**	0.5050 (0.0036)
0.25	35	150	**0.2349 (0.0045)**	**0.2396 (0.0046)**	0.2517 (0.0054)	**0.2664 (0.0066)**
0.25	35	500	**0.2267 (0.0023)**	**0.2318 (0.0024)**	**0.2409 (0.0026)**	0.2513 (0.0030)
0.15	35	500	**0.1379 (0.0020)**	**0.1406 (0.0021)**	**0.1464 (0.0025)**	0.1541 (0.0029)

Table 5.6: Average of Estimates for the Mean Age of Onset (Generated by Logistic)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	79.624 (0.218)	79.880 (0.231)	**80.634 (0.269)**	**81.464 (0.315)**
0.5	80	500	79.879 (0.104)	**80.328 (0.120)**	**81.518 (0.183)**	**82.490 (0.224)**
0.25	80	150	79.690 (0.284)	79.793 (0.303)	**80.775 (0.377)**	**81.857 (0.497)**
0.25	80	500	79.999 (0.160)	**80.408 (0.188)**	**81.523 (0.221)**	**82.635 (0.307)**
0.15	80	500	79.935 (0.211)	80.518 (0.259)	**81.935 (0.466)**	**83.461 (0.656)**
0.5	55	150	55.103 (0.101)	55.106 (0.102)	**56.149 (0.150)**	**57.269 (0.323)**
0.5	55	500	54.950 (0.052)	54.986 (0.055)	**55.910 (0.092)**	**56.826 (0.147)**
0.25	55	150	54.884 (0.175)	54.963 (0.182)	**55.918 (0.260)**	**57.184 (0.715)**
0.25	55	500	55.019 (0.088)	55.064 (0.091)	**55.952 (0.118)**	**57.080 (0.214)**
0.15	55	500	54.932 (0.098)	54.923 (0.101)	**55.872 (0.156)**	**56.975 (0.220)**
0.5	35	150	34.902 (0.176)	**35.436 (0.190)**	**39.457 (0.385)**	**43.636 (0.798)**
0.5	35	500	34.990 (0.078)	**35.637 (0.091)**	**43.185 (0.533)**	**49.081 (0.747)**
0.25	35	150	34.813 (0.257)	**35.371 (0.295)**	**41.219 (0.762)**	**45.289 (1.039)**
0.25	35	500	34.877 (0.129)	**35.539 (0.158)**	**44.367 (0.848)**	**56.974 (1.652)**
0.15	35	500	35.019 (0.154)	**35.699 (0.191)**	**45.355 (1.131)**	**58.143 (2.539)**

Table 5.7: Average of Estimates for Mean Age of Onset (Generated by Normal)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	79.692 (0.214)	79.953 (0.226)	80.417 (0.241)	**81.156 (0.287)**
0.5	80	500	**79.709 (0.118)**	79.981 (0.120)	**80.652 (0.136)**	**81.168 (0.147)**
0.25	80	150	**79.094 (0.320)**	**79.158 (0.328)**	80.001 (0.386)	80.498 (0.462)
0.25	80	500	**79.721 (0.167)**	79.949 (0.171)	**80.605 (0.192)**	**81.166 (0.213)**
0.15	80	500	79.845 (0.217)	79.951 (0.223)	**80.631 (0.256)**	**81.202 (0.276)**
0.5	55	150	55.060 (0.109)	55.052 (0.107)	**55.689 (0.130)**	**56.464 (0.224)**
0.5	55	500	**54.873 (0.060)**	54.899 (0.058)	**55.443 (0.061)**	**56.038 (0.074)**
0.25	55	150	55.184 (0.169)	55.173 (0.167)	**55.851 (0.196)**	**56.695 (0.324)**
0.25	55	500	54.945 (0.093)	54.955 (0.092)	**55.547 (0.106)**	**56.141 (0.120)**
0.15	55	500	55.065 (0.125)	55.123 (0.122)	**55.718 (0.145)**	**56.398 (0.176)**
0.5	35	150	**34.518 (0.172)**	34.891 (0.186)	**38.350 (0.359)**	**43.353 (0.657)**
0.5	35	500	**34.460 (0.087)**	34.913 (0.095)	**40.364 (0.317)**	**48.384 (0.617)**
0.25	35	150	**34.394 (0.251)**	34.874 (0.290)	**40.515 (0.773)**	**46.699 (1.618)**
0.25	35	500	**34.704 (0.139)**	35.168 (0.153)	**41.810 (0.596)**	**51.305 (0.987)**
0.15	35	500	**34.322 (0.171)**	34.711 (0.182)	**41.872 (0.745)**	**52.721 (1.901)**

Table 5.8: Average of Estimates for Mean Age of Onset (Generated by Gamma)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	**79.346 (0.212)**	79.648 (0.217)	80.085 (0.234)	80.491 (0.252)
0.5	80	500	**79.263 (0.106)**	**79.575 (0.111)**	80.044 (0.120)	**80.406 (0.128)**
0.25	80	150	**79.032 (0.329)**	**79.090 (0.326)**	79.635 (0.347)	79.796 (0.371)
0.25	80	500	**79.189 (0.166)**	**79.477 (0.170)**	79.880 (0.181)	80.259 (0.193)
0.15	80	500	79.630 (0.228)	79.848 (0.225)	80.349 (0.242)	**80.688 (0.260)**
0.5	55	150	**53.982 (0.118)**	**54.203 (0.115)**	54.554 (0.122)	**54.886 (0.130)**
0.5	55	500	**53.892 (0.060)**	**54.154 (0.060)**	54.476 (0.062)	**54.780 (0.066)**
0.25	55	150	**54.057 (0.152)**	54.262 (0.155)	54.591 (0.166)	**54.914 (0.180)**
0.25	55	500	**53.919 (0.094)**	**54.176 (0.095)**	54.491 (0.099)	**54.793 (0.105)**
0.15	55	500	**53.885 (0.121)**	**54.179 (0.121)**	54.480 (0.127)	54.772 (0.135)
0.5	35	150	**32.417 (0.142)**	**32.712 (0.145)**	34.282 (0.210)	**36.403 (0.339)**
0.5	35	500	**32.405 (0.074)**	**32.745 (0.077)**	34.161 (0.107)	**35.876 (0.154)**
0.25	35	150	**32.057 (0.202)**	**32.307 (0.202)**	33.935 (0.339)	**35.898 (0.499)**
0.25	35	500	**32.497 (0.111)**	**32.872 (0.119)**	34.412 (0.161)	**36.464 (0.241)**
0.15	35	500	**32.421 (0.147)**	**32.828 (0.158)**	34.621 (0.242)	**37.211 (0.488)**

Table 5.9: Average of Estimates for Mean Age of Onset (Generated by Lognormal)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	**79.312 (0.209)**	**79.481 (0.207)**	79.834 (0.225)	80.198 (0.245)
0.5	80	500	**78.950 (0.113)**	**79.322 (0.118)**	**79.670 (0.126)**	79.941 (0.133)
0.25	80	150	**78.587 (0.337)**	**78.729 (0.342)**	79.408 (0.373)	79.356 (0.391)
0.25	80	500	**79.088 (0.166)**	**79.367 (0.170)**	79.736 (0.185)	80.020 (0.191)
0.15	80	500	**78.918 (0.229)**	**79.294 (0.250)**	**79.456 (0.239)**	79.942 (0.281)
0.5	55	150	**54.002 (0.102)**	**54.342 (0.105)**	**54.605 (0.112)**	54.805 (0.116)
0.5	55	500	**54.228 (0.059)**	**54.604 (0.060)**	**54.852 (0.063)**	55.073 (0.064)
0.25	55	150	**54.411 (0.153)**	**54.641 (0.154)**	54.959 (0.164)	55.195 (0.173)
0.25	55	500	**54.198 (0.088)**	**54.602 (0.091)**	54.832 (0.098)	55.070 (0.098)
0.15	55	500	**54.218 (0.114)**	**54.567 (0.117)**	54.811 (0.124)	55.048 (0.127)
0.5	35	150	**32.959 (0.162)**	**33.372 (0.171)**	**34.448 (0.216)**	**35.586 (0.278)**
0.5	35	500	**32.722 (0.076)**	**33.173 (0.082)**	**34.134 (0.101)**	35.167 (0.127)
0.25	35	150	**32.557 (0.189)**	**32.946 (0.196)**	**34.034 (0.261)**	35.474 (0.383)
0.25	35	500	**32.601 (0.105)**	**33.010 (0.110)**	**33.962 (0.138)**	35.087 (0.180)
0.15	35	500	**32.619 (0.140)**	**32.962 (0.145)**	**33.917 (0.190)**	35.199 (0.257)

Table 5.10: Average of Estimates for the Standard Deviation of the Age of Onset (Generated by Logistic)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	**9.564 (0.132)**	9.736 (0.153)	**10.633 (0.207)**	**11.617 (0.268)**
0.5	80	500	9.835 (0.075)	**10.193 (0.092)**	**11.522 (0.150)**	**12.619 (0.194)**
0.25	80	150	**9.196 (0.199)**	**9.249 (0.222)**	10.388 (0.301)	**11.510 (0.469)**
0.25	80	500	**9.751 (0.105)**	10.062 (0.126)	**11.325 (0.186)**	**12.503 (0.263)**
0.15	80	500	**9.565 (0.146)**	9.927 (0.191)	**11.263 (0.338)**	**12.750 (0.509)**
0.5	55	150	9.892 (0.085)	**9.779 (0.089)**	**11.294 (0.197)**	**12.742 (0.386)**
0.5	55	500	9.989 (0.051)	**9.883 (0.053)**	**11.321 (0.119)**	**12.649 (0.185)**
0.25	55	150	9.878 (0.139)	**9.642 (0.140)**	**10.902 (0.262)**	**12.612 (0.820)**
0.25	55	500	10.003 (0.069)	9.879 (0.074)	**11.206 (0.134)**	**12.775 (0.258)**
0.15	55	500	9.984 (0.084)	9.862 (0.084)	**11.259 (0.182)**	**12.771 (0.271)**
0.5	35	150	9.826 (0.116)	**9.919 (0.122)**	**13.834 (0.358)**	**18.916 (0.902)**
0.5	35	500	9.952 (0.059)	10.125 (0.063)	**16.747 (0.428)**	**23.941 (0.762)**
0.25	35	150	**9.567 (0.145)**	**9.639 (0.160)**	**14.696 (0.576)**	**20.486 (1.082)**
0.25	35	500	9.856 (0.084)	10.049 (0.094)	**17.528 (0.652)**	**32.297 (1.809)**
0.15	35	500	9.891 (0.111)	10.063 (0.126)	**17.821 (0.872)**	**33.896 (2.972)**

Table 5.11: Average of Estimates for Standard Deviation of the Age of Onset (Generated by Normal)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	**9.599 (0.128)**	**9.551 (0.137)**	10.245 (0.168)	**11.075 (0.225)**
0.5	80	500	**9.845 (0.070)**	**9.787 (0.071)**	**10.632 (0.093)**	**11.321 (0.111)**
0.25	80	150	**8.857 (0.192)**	**8.690 (0.197)**	9.663 (0.252)	10.185 (0.334)
0.25	80	500	**9.729 (0.101)**	**9.653 (0.103)**	**10.480 (0.132)**	**11.200 (0.160)**
0.15	80	500	**9.600 (0.123)**	**9.497 (0.123)**	10.308 (0.164)	**11.058 (0.196)**
0.5	55	150	**10.303 (0.089)**	9.859 (0.084)	**10.859 (0.134)**	**11.951 (0.265)**
0.5	55	500	**10.404 (0.046)**	9.963 (0.041)	**10.860 (0.056)**	**11.784 (0.079)**
0.25	55	150	10.227 (0.131)	9.791 (0.124)	**10.790 (0.182)**	**11.987 (0.365)**
0.25	55	500	**10.293 (0.066)**	9.883 (0.063)	**10.804 (0.089)**	**11.744 (0.117)**
0.15	55	500	**10.370 (0.090)**	9.935 (0.087)	**10.826 (0.125)**	**11.864 (0.179)**
0.5	35	150	9.967 (0.111)	**9.753 (0.114)**	**13.190 (0.276)**	**19.336 (0.720)**
0.5	35	500	**10.137 (0.056)**	9.956 (0.057)	**15.080 (0.264)**	**24.312 (0.677)**
0.25	35	150	**9.712 (0.141)**	**9.612 (0.156)**	**14.754 (0.589)**	**23.248 (1.963)**
0.25	35	500	10.088 (0.087)	9.933 (0.090)	**15.824 (0.454)**	**26.619 (0.970)**
0.15	35	500	9.952 (0.109)	**9.759 (0.110)**	**15.934 (0.573)**	**29.452 (2.358)**

Table 5.12: Average of Estimates for Standard Deviation of the Age of Onset (Generated by Gamma)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	**9.198 (0.121)**	**9.049 (0.122)**	**9.677 (0.147)**	10.176 (0.170)
0.5	80	500	**9.445 (0.067)**	**9.316 (0.068)**	9.979 (0.081)	**10.472 (0.094)**
0.25	80	150	**8.673 (0.179)**	**8.349 (0.172)**	**9.031 (0.201)**	**9.308 (0.234)**
0.25	80	500	**9.217 (0.098)**	**9.087 (0.095)**	**9.698 (0.113)**	10.187 (0.133)
0.15	80	500	**9.399 (0.126)**	**9.205 (0.123)**	9.882 (0.148)	**10.366 (0.170)**
0.5	55	150	**9.865 (0.086)**	**9.375 (0.079)**	9.921 (0.093)	**10.479 (0.111)**
0.5	55	500	**9.878 (0.042)**	**9.444 (0.039)**	9.965 (0.045)	**10.499 (0.053)**
0.25	55	150	**9.656 (0.116)**	**9.175 (0.113)**	9.701 (0.132)	**10.245 (0.159)**
0.25	55	500	**9.755 (0.063)**	**9.318 (0.058)**	9.832 (0.068)	**10.356 (0.080)**
0.15	55	500	**9.742 (0.088)**	**9.304 (0.081)**	9.792 (0.094)	**10.302 (0.110)**
0.5	35	150	**8.471 (0.090)**	**8.097 (0.084)**	9.808 (0.153)	**12.331 (0.302)**
0.5	35	500	**8.471 (0.049)**	**8.111 (0.047)**	**9.690 (0.079)**	**11.795 (0.137)**
0.25	35	150	**8.220 (0.123)**	**7.796 (0.118)**	9.470 (0.232)	**11.911 (0.458)**
0.25	35	500	**8.584 (0.068)**	**8.223 (0.066)**	9.920 (0.114)	**12.364 (0.212)**
0.15	35	500	**8.577 (0.083)**	**8.242 (0.085)**	10.134 (0.162)	**13.139 (0.433)**

Table 5.13: Average of Estimates for the Standard Deviation of the Age of Onset (Generated by Lognormal)

p	mean	Relatives	Logistic	Normal	Gamma	Lognormal
0.5	80	150	**8.947 (0.135)**	**8.732 (0.131)**	**9.244 (0.154)**	9.715 (0.180)
0.5	80	500	**9.117 (0.072)**	**8.982 (0.071)**	**9.515 (0.083)**	9.911 (0.093)
0.25	80	150	**8.240 (0.202)**	**7.960 (0.199)**	**8.764 (0.231)**	**8.838 (0.265)**
0.25	80	500	**9.072 (0.098)**	**8.908 (0.096)**	**9.454 (0.115)**	9.877 (0.127)
0.15	80	500	**8.974 (0.118)**	**8.829 (0.123)**	**9.282 (0.134)**	9.786 (0.165)
0.5	55	150	**9.320 (0.082)**	**8.932 (0.078)**	**9.339 (0.089)**	**9.724 (0.100)**
0.5	55	500	**9.638 (0.044)**	**9.220 (0.041)**	**9.596 (0.045)**	10.015 (0.051)
0.25	55	150	**9.262 (0.115)**	**8.824 (0.107)**	**9.305 (0.125)**	9.745 (0.146)
0.25	55	500	**9.579 (0.066)**	**9.190 (0.063)**	**9.575 (0.071)**	9.994 (0.077)
0.15	55	500	**9.539 (0.082)**	**9.104 (0.078)**	**9.492 (0.084)**	9.932 (0.100)
0.5	35	150	**7.899 (0.094)**	**7.596 (0.094)**	**8.798 (0.139)**	10.235 (0.210)
0.5	35	500	**7.950 (0.048)**	**7.646 (0.048)**	**8.748 (0.070)**	10.073 (0.103)
0.25	35	150	**7.698 (0.120)**	**7.395 (0.117)**	**8.591 (0.183)**	10.298 (0.322)
0.25	35	500	**7.884 (0.071)**	**7.566 (0.067)**	**8.678 (0.099)**	10.080 (0.150)
0.15	35	500	**7.715 (0.086)**	**7.390 (0.083)**	**8.507 (0.131)**	10.025 (0.205)

6 CONCLUSIONS

6.1 Risks for Alzheimer's and Their Implications

The three general procedures of estimation all suggest that the lifetime risk for disease does not approach 50%. In particular the risk of dementia by age 90 appears to be approximately only 25%, much lower than the 50% risk found in other recent studies. Three possibilities for the differences between this data set and others come to mind. The stringent criterion with its requirement of existence of medical records appears to be much stronger than the criteria used in the other studies mentioned. The relaxed criteria still may be more restrictive than that used by other studies. In the product-limit setting, Breitner and Magruder-Habib (1989) showed that varying the rule for deciding age of onset could change risk estimates. As the age-specific risk estimates from the full maximum likelihood procedure appear to be close to the product-limit estimates, changing the rule determining the age of onset should also affect the full maximum likelihood estimates. Possibly the age of onset has been determined differently in some of the other studies. Finally, as the other groups ascertain their index cases differently, the populations being sampled may be different. As the other studies were conducted in the United States, one factor which could affect sampling is the universal medical insurance program available in Canada. Also the methods of ascertainment used by other groups may cause more families with the genetic form of the disease to be included in the sample.

6.2 Comparisons of the Estimation Methods

Three types of estimation procedures for calculating risks for disease have been discussed. The product-limit method gives a very easy and quick way to estimate age-specific risks. The Kaplan-Meier and life-table estimators appear to be approximately unbiased, while the Weinberg estimator appears to have a positive bias, suggesting that it shouldn't be used. The one drawback to these three estimation procedures is that it is difficult to estimate the lifetime risk for disease without making possibly unreasonable

assumptions. The estimators for lifetime risk using fixed approximations to the age of onset distributions, while easy to calculate, do not appear to be very useful. It appears that they can be strongly influenced by the choice for the age of onset distribution, with little robustness when a poor choice is made. Also, one estimator, the modified Strömgren, will almost always be biased, even when the correct age of onset distribution is chosen. The maximum likelihood procedure for estimating the lifetime risk and the age of onset distribution appears to be the most useful. Since the age of onset distribution is estimated along with the lifetime risk, the lifetime risk estimate should have a smaller bias than the fixed onset distribution estimators, though poor choices for the class of the age of onset distributions could still lead to biased estimates. However, the age-specific risks calculated by this method appear to be very close to those calculated by Kaplan-Meier or the life-table estimators, irrespective of the choice of the class of the age of onset distributions. Our simulation study seems to support the usefulness of this type of estimator when intelligent choices are made about which classes of distributions to use to estimate the age of onset.

BIBLIOGRAPHY

1 Anscombe, F.J. (1964) Normal Likelihood Functions. Ann. Inst. Stat. Math (Tokyo), 16:1-19.

2 Armenian, H.K., Khoury, M.J. (1981) Age at Onset of Genetic Diseases: an Application for Sartwell's Model of the Distribution of Incubation Periods. Am. J. Epidemiol., 113:596-605.

3 Breitner, J.C.S., Folstein, M.F. (1984) Familial Alzheimer Dementia: A Prevalent Disorder with Specific Clinical Features. Psychol. Med., 14:63-80.

4 Breitner, J.C.S., Folstein, M.F., Murphy, E.A. (1986) Familial Aggregation in Alzheimer Dementia–I. A Model for the Age-dependent Expression of an Autosomal Dominant Gene. J. Psychiat. Res., 20:31-43.

5 Breitner, J.C.S., Magruder-Habib, K.M. (1989) Criteria for Onset Critically Influence the Estimation of Familial Risk in Alzheimer's Disease. Submitted, Genet Epidemiol.

6 Breitner, J.C.S., Murphy, E.A., Silverman, J.M., et al. (1988a) Age-Dependent Expression of Familial Risk in Alzheimer's Disease. Am. J. Epidemiol, 128:536-548.

7 Breitner, J.C.S., Silverman, J.M., Mohs, R.C., Davis, J.L. (1988b) Familial Aggregation in Alzheimer's Disease: Comparison of Risk Among Relatives of Early- and Late-Onset Cases, and Among Male and Female Relatives in Successive Generations. Neurology, 38:207-212.

8 Chase, G.A., Folstein, M.F., Breitner, J.C.S., et al. (1983) The Use of Life Tables and Survival Analysis in Testing Genetic Hypotheses, With an Application to Alzheimer's Disease. Am. J. Epidemiol., 117:590-597.

9 Editorial. (1986) Researchers Hunt for Alzheimer's Disease Gene. Science, 232:448-450.

Farrer, L.A., O'Sullivan, D.M., Cupples, L.A., et al. (1989) Assessment of Genetic Risk for Alzheimer's Disease Among First-Degree Relatives, Ann. Neurol., 25:485-493.

10 Friedland, R.P. (Moderator). (1988) NIH Conference. Alzheimer's Disease: Clinical and Biological Heterogeneity. Ann. Int. Med., 109:298-311.

11 Greenwood, M. (1926) The Natural Duration of Cancer. Reports of Public Health and Medical Subjects, Vol 33. London: Her Majesty's Stationary Office.

12 Horner, R.D. (1987) Age at Onset of Alzheimer's Disease: Clue to the Relative Importance of Etiologic Factors? Am. J. Epidemiol., 126:409-414.

13 Joachim, C.L., Morris, J.H., Selkoe, D.J. (1988) Clinically Diagnosed Alzheimer's Disease: Autopsy Results in 150 Cases. Ann. Neurol., 24:50-56.

14 Kaplan, E.L., Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. JASA, 58:457-481.

15 Katzman R. (1976) The Prevalence and Malignancy of Alzheimer's disease: A Major Killer. Arch. Neurol., 33:217-218.

16 Larsson, T., Sjögren, T. (1954) A Methological, Psychiatric and Statistical Study of a Large Swedish Rural Population. Acta Psychiatrica et Neurological Scandinavica (Supplement), 89:40-54.

17 Lawless, J.F. (1982) Statistical Models and Methods for Lifetime Data. New York: John Wiley and Sons.

18 McKhann, G., Drachman, D., Folstein, M., et al. (1984) Clinical Diagnosis of Alzheimer's Disease: Report of the NINCDS-ADRDA Work Group Under the Auspices of the Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology, 34:939-944.

19 Marsden C.D. (1978) The Diagnosis of Dementia. In Studies in Geriatric Psychiatry. Issacs AD, Post F. (eds). New York: John Wiley and Sons, 99-110.

20 Martin, R.L., Gerteis, G., Gabrielli, W.F. (1988) A Family-Genetic Study of Dementia of the Alzheimer Type. Arch. Gen. Psychiatry, 45:894-900.

21 Marx, J.L. (1988) Evidence Uncovered for a Second Alzheimer's Gene. Science, 241:1432-1433.

22 Morales, A.J., Murphy, E.A., Krush, A.J. (1984) The Bingo Model of Survivorship. II: Statistical Aspects of the Bingo Model of Multiplicity 1 With Application to Heredity Polyposis of the Colon. Am. J. Med. Genet., 17:783-801.

23 Pericak-Vance, M.A., Elston, R.C., Conneally, P.M., et al. (1983) Age-of-Onset Heterogeneity in Huntington Disease Families. Am. J. Med. Genet., 14:49-59.

24 Pericak-Vance, M.A., Yamaoka, L.H., Haynes, C.S., et al. (1988) Genetic Linkage Studies in Alzheimer's Disease Families. Exp. Neurol., 102:271-279.

25 Risch, N. (1983) Estimating Morbidity Risks with Variable Age of Onset: Review of Methods and a Maximum Likelihood Approach. Biometrics, 39:929-939.

26 Sadovnick, A.D., Baird, P.A. (1988) The Familial Nature of Multiple Sclerosis: Age-corrected Empiric Recurrence Risks for Parents and Siblings of Patients. Neurology, 38:990-991.

27 Sadovnick, A.D., Irwin, M.E., Baird, P.A., et al. (1989) Genetic Studies on an Alzheimer Clinic Population. Genet. Epidemiol. (In Press).

28 Sadovnick, A.D., Tuokko, H., Horton, et al. (1988) Familial Alzheimer's Disease. Can. J. Neurol. Sci., 15:142-146.

29 St. George-Hyslop, P.H., Tanzi, R.E., Polinsky, R.J., et al. (1987) The Genetic Defect Causing Familial Alzheimer's Disease Maps on Chromosome 21. Science, 235:885-890.

30 Sartwell, P.E., The Distribution of Incubation Periods of Infectious Disease. (1950) Am. J. Publ. Health, 51:310-318.

31 Schellenberg, G.D., Bird, T.D., Wijsman, E.M., et al. (1988) Absence of Linkage of Chromosome 21q21 Markers to Familial Alzheimer's Disease. Science, 241:1507-1510.

32 Schulz, B. (1937) Ubersicht uber auslesefreie Untersuchungen in der verwandtschaft Manisch-depressiven. Zeitschrift fur Psychische Hygiene, 10:39-60.

33 Slater, E., Cowie, V. (1971) The Genetics of Mental Disorders. London, Oxford University Press.

34 Strömgren, E. (1935) Zum ersatz des Weinbergschen 'abgekurzten verfahrens' zugleich ein beitrag zur Frage von der Erblichkeit des Erkrankungsalters bei der Schizophrenie. Zeitschrift fur die gesamte Neurologie und Psychatrie, 1935;153:784-797.

35 Strömgren, E. (1938) Beitrage zur psychiatrischen erblehre auf grund von Untersuchungen an einer Inselbevolkerung. Acta Psychiatrica et Neurologica Scandinavica (Supplement), 19:1-257.

36 Thompson, W.D., Weissman, M.M. (1981) Quantifying Lifetime Risk of Psychiatric Disorder. J. Psychiat. Res, 16:113-126.

37 Tierney, M.C., Fisher, R.H., Lewis, A.J., et al. (1988) The NINCDS-ADRDA Workgroup Criteria for the Clinical Diagnosis of Alzheimer's Disease: A Clinico-pathologic Study of 57 Cases. Neurology, 38:359-364.

38 Weinberg, W. (1925) Methoden und Tecknik der Statistik mit besonderer berucksichtigung der Sozialbiologie. In Handbuch der Sozialen Hygiene und Gesundheitsfursorge, Gottstein, A., Schlossman, A., Telely, L. (eds), Berlin.

39 Winokur, G., Clayton, P.J., Reich, T. (1969) Manic-Depressive Illness. St. Louis: Mosby.

40 Zubenko, G.S., Huff, F.J., Beyer, J., et al. (1988) Familial Risk of Dementia Associated with a Biologic Subtype of Alzheimer's Disease, 45:889-893.