Computational and Mathematical Methods in Medicine
Vol. 11, No. 2, June 2010, 119–159

Taylor & Francis
Taylor & Francis Group

Calculating the number of people with Alzheimer's disease in any country using saturated mutation models of brain cell loss that also predict widespread natural immunity to the disease

Ivan Kramer*

Department of Physics, University of Maryland Baltimore County, 1000 Hilltop Circle, Catonsville, MD 21250, USA

(*Received 3 May 2007; final version received 16 March 2009*)

The series of mutations that cause brain cells to spontaneously and randomly die leading to Alzheimer's disease (AD) is modelled. The prevalence of AD as a function of age in males and females is calculated from two very different mutation models of brain cell death. Once the prevalence functions are determined, the number of people with AD in any country or city can be estimated.

The models developed here depend on three independent parameters: the number of mutations necessary for a brain cell associated with AD to spontaneously die, the average time between mutations, and the fraction of the risk population that is immune to developing the disease, if any. The values of these parameters are determined by fitting the model's AD incidence function to the incidence data.

The best fits to the incidence rate data predict that as much as 74.1% of males and 79.5% of females may be *naturally immune* to developing AD. Thus, the development of AD is *not a normal or inevitable* result of the aging process. These fits also predict that males and females develop AD through different pathways, requiring a different number of mutations to cause the disease. The number of people in the USA with AD in the year 2000 is estimated to be 451,000.

It is of paramount importance to determine the nature of the immunity to AD predicted here. Finding ways of blocking the mutations leading to the random, spontaneous death of memory brain cells would prevent AD from developing altogether.

Keywords: Alzheimer's disease; mutation model; prevalence; incidence rate; brain cell loss

1. The ordered mutation model of late-onset Alzheimer's disease

Alzheimer's disease (AD) begins when neurons in the brain spontaneously begin dying in a disorganized way taking some memory with it. Research data suggests that AD results from a series of *random* missense mutations in brain cells. In this section, the possibility that *all* these mutations occur one-by-one *in a definite order* in a neuron until the final mutation causes the cell to die will be explored. For example, in this *ordered* mutation model a normal, unmutated brain all can only mutate into the first mutation state *and no other*, and cells that are in the first mutation state can only mutate into the second mutation state *and no other*, etc. It is assumed that the characteristics of the cell in each mutation state are unique and different from the characteristics of the cell in every other mutation

*Email: kramer@umbc.edu

ISSN 1748-670X print/ISSN 1748-6718 online

DOI: 10.1080/17486700902910076
http://www.informaworld.com

state. It is further assumed that AD progresses as more and more neurons die. At some point, the disease impacts the central nervous system to such an extent that the patient dies.

Clearly, the number of mutations in a brain cell necessary to cause the onset of clinical symptoms of AD in males or females (risk populations) is important to determine. Equally important to ascertain is the average time a cell spends in any particular mutation state, a quantity that will be called the mutation *lifetime* of the state. Perhaps the most important thing to calculate from modelling the AD incidence rate in a particular risk population is the prevalence of natural immunity to AD if, indeed, any exists. The model to be presented below will enable the values of all three of these quantities to be calculated by fitting the model's incidence function to AD incidence data.

An important, novel feature of the AD model constructed here is that it is inherently saturated, *i.e.* the maximum percentage of a risk population that can develop AD can never exceed 100%. Thus, the model's AD incidence rate function always increases, peaks, and subsequently declines towards zero as the age of the risk population increases to arbitrarily large values (in practice, the peak could lie *above* the maximum human lifespan).

In this model, it will be assumed that an ordered series of mutations, numbering m, of a brain cell is required for the cell to spontaneously die. At any age t, every brain cell of a person is in one of the mutation states. *The probability or risk of a person developing AD is exactly equal to the probability that a random brain cell was mutated to death in this fashion.*

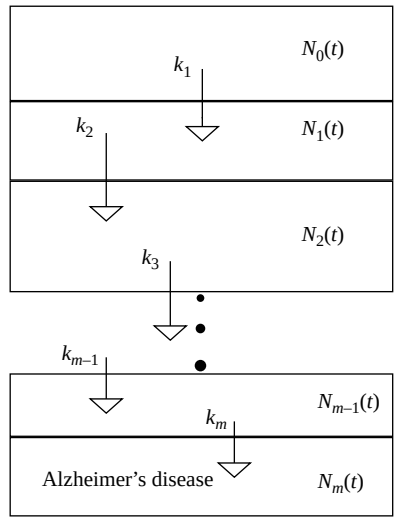
Suppose the *total* number of brain cells in the average member of a risk population is denoted by N_T. If a fraction f_s of these cells are susceptible to the ordered mutations leading to death of this model, then the number of susceptible brain cells in the average brain of the risk population is $N_s = f_s N_T$.

Consider a random representative cohort of a risk population whose members all have the same age t. Here, $t = 0$ is coincident with birth. It will also be assumed that a total of m mutations are required in a brain cell for it to spontaneously die.

Then, the number of brain cells in the average brain of the cohort that are in the rth mutation state at age t will be denoted by $N(r/m,t)$ or $N_r(t)$ for short, where $r = 0,1,2,\ldots,m$. Assuming that all cells in the average brain are in the zeroth mutation state at birth, then at age $t = 0, N_r(0) = 0$ for $r = 1,2,3,\ldots,m$, and $N_0(0) = N_s$. A schematic representation of the values of the $N_p(t)$ at age t is shown in Figure 1 where the ordered mutations are represented by the arrows connecting two sequential mutation states.

Consider the first mutation of a brain cell necessary to cause AD. The fraction of an average brain with no mutated cells that experiences the first mutation per unit time will be denoted by k_1 and called the first *mutation rate*. The average time required for a brain cell to experience the first mutation will be defined as the first *mutation lifetime* and will be denoted by T_1. The mutation lifetime and the mutation rate will be shown to be reciprocals of each other so that $k_1 \equiv T_1^{-1}$. Continuing in this way, the number of cells in the average brain $N_r(t)$ who are in the rth mutation state at age t depends on the values of the r mutation rates $(k_r) \equiv [k_1, k_2, \ldots, k_r]$ as depicted in Figure 1. It will be assumed that all the mutation rates are constants so that the mutations experienced by the cohort *occur randomly*.

The Appendix contains a complete description on how the functions $N_r(t)$ are computed from the model just described. One of the most important features of the model is that it allows the possibility that a fraction f_i of the brain cells are *naturally immune* to the mutation process leading to death described here. The fraction f_s of brain cells that are susceptible to mutation to death is related to the fraction that is immune to it since $f_s + f_i = 1$.



$$N_s = N_0(t) + N_1(t) + N_2(t) + \ldots + N_{m-1}(t) + N_m(t)$$

Figure 1. Mutation model of brain cells leading to AD.

The probability $P(t)$ that a neuron will undergo the ordered set of m mutations and die at age t is given by

$$P(t) = \frac{N_m(t)}{N_T}. \tag{1}$$

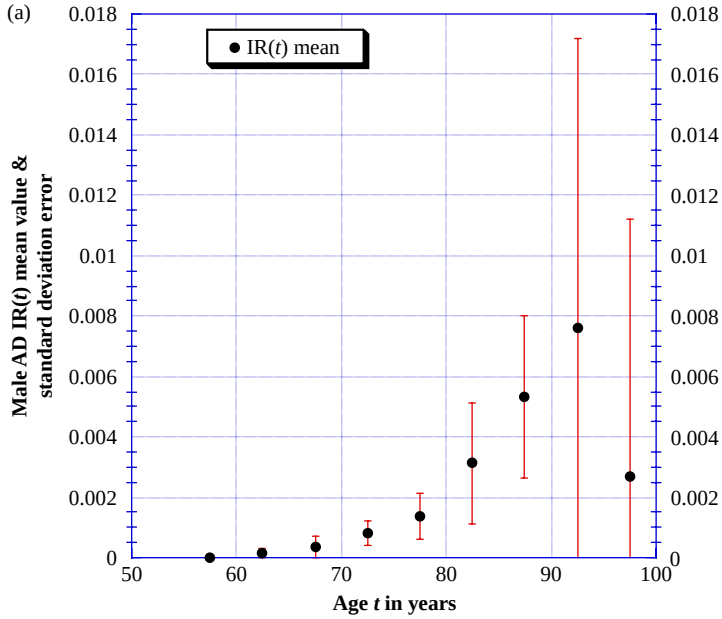


Figure 1a. Male AD data for annual incidence rate IR(t) mean value and standard deviation error as a function of age t in years from Table 1(c).

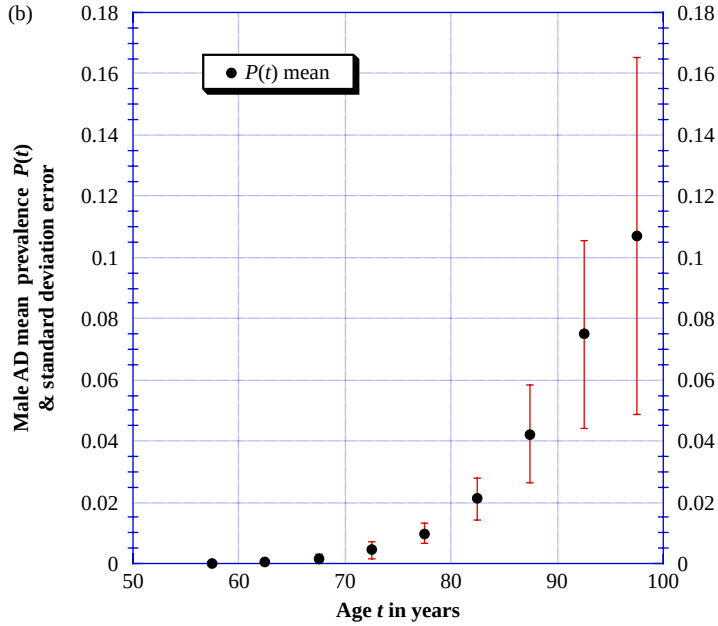


Figure 1b. Male AD data for mean prevalence $P(t)$ and standard deviation error as a function of age t in years from Table 1 (b2).

This probability is also equal to the probability of developing AD at age t in the risk population, a quantity also known as the prevalence of AD at age t. Thus, in a population of a total number of H_{T} humans, if the number of humans that have been diagnosed with AD by age t is denoted by $H_{\mathrm{AD}}(t)$, then

$$P(t) = \frac{H_{\mathrm{AD}}(t)}{H_{\mathrm{T}}} = \frac{N_m(t)}{N_{\mathrm{T}}}. \tag{2}$$

The fraction of the population that comes down with AD per unit time at age t (the *fractional AD incidence rate*) is given by

$$\mathrm{IR}(m;(k_m);t) \equiv \frac{1}{H_{\mathrm{T}}}\frac{\mathrm{d}H_{\mathrm{AD}}(t)}{\mathrm{d}t} = \frac{\mathrm{d}P(t)}{\mathrm{d}t}, \tag{3}$$

where m is the number of mutations necessary to cause an average brain cell to spontaneously die, (k_m) is the set of m mutation rates, and f_{s} is the fraction of brain cells that are susceptible to mutation to death. All these model parameters are determined by fitting the incident rate function to AD incidence data.

The AD incidence rate $\mathrm{IR}(t)$ and prevalence $P(t)$ functions arising from the ordered mutation model described here are derived in the Appendix. For the particularly simple case when all the mutation rates are equal to each other, the AD incidence rate is given by the particular simple function given in (A8) in the Appendix; integrating this function gives the AD prevalence function $P(t)$ given in (A7). Notice that as the age t of the cohort becomes arbitrarily large, the prevalence function in (A7) approaches f_{s} in value. Since $f_{\mathrm{s}} \leq 1$, the prevalence function is *naturally saturated* in that it can never exceed unity in value. The incidence rate function given in (A7) characteristically monotonically rises

to a peak as the age t of the cohort increases from birth and then monotonically decreases to zero as t continues increasing. Thus, the area under the IR(t) curve, namely $P(t)$, is finite and can never exceed f_s in value. These particularly simple functions will be shown to give surprisingly good fits to the AD data.

Fitting the incidence rate function IR(t) to the AD incidence rate data determines the values of the model parameters. If every member of the population is susceptible to developing AD, then $f_s = 1$ and the fit involves determining the values of the remaining model parameters.

The *saturated ordered* mutation model constructed and solved in this paper is isomorphic to the physical model describing an ordered chain of radioactive nuclei decays with the exception that it allows for the possibility that a fraction of a risk population may be *immune* to developing AD.

2. The *un*ordered, *independent* mutation model of AD

A completely different model of the mutations in a brain cell that lead to AD postulates that every mutation occurs *independently* of all the others. In Figure 1, only the first transition is permitted for each mutation and from Equation (A3a) in the Appendix with $f_s = 1$, the fraction of brain cells that have made the first mutation necessary to cause AD as a function of age is

$$p_1(t) \equiv 1 - \exp(-k_1 t), \tag{4}$$

where the transition constant k_1 is related to average time T_1 necessary for this mutation to occur in a cell by $T_1 = 1/k_1$. If m *independent* mutations are required for a brain cell to be destroyed, then the probability that a cell will be destroyed at age t is given by

$$P(t) = p_1(t)p_2(t)p_3(t)\ldots p_{m-1}(t)p_m(t), \tag{5}$$

where the values of the m mutation constants $k_1, k_2, \ldots, k_{m-1}, k_n$ are all independent of each other in general. Notice that the mutations here are not ordered, as they where in the model described in section 1 above, but are completely independent of each other. Thus, in this model the mutations can occur in any order, simultaneously, or at completely different times.

In the limit when $k_i t \ll 1$ for all i, (5) gives

$$P(t) = k_1 k_2 k_3 \ldots k_m \times t^m, \quad k_i t \ll 1. \tag{6}$$

This result is to be compared with (A4) in the Appendix obtained from the ordered model.

The simplest such independent mutation model occurs, if all the mutation constants have the same positive value k. Under these circumstances, using Equations (4) and (5), the probability that a brain cell will be destroyed at time t becomes

$$P(t) = [1 - \exp(-kt)]^m. \tag{7}$$

The quantity k defined in (7) can be regarded as an *effective* mutation constant that reproduces the experimental value of $P(t)$ in a simplified simulation of the data.

In the general case, where only a fraction f_s of the population is susceptible to developing AD, the prevalence of AD at age t is then given by

$$P(t) = f_s[1 - \exp(-kt)]^m. \tag{8}$$

Since $P(t)$ also represents the probability of developing AD at age t in the population, the incidence rate of AD at age t is given by

$$\mathrm{IR}(t) = \frac{\mathrm{d}P(t)}{\mathrm{d}t} \equiv I'(f_s, k, m) = f_s mk[1 - \exp(-kt)]^{(m-1)}\exp(-kt). \tag{9}$$

Thus, in a simplified simulation of AD incidence rate data using the independent mutation model, the values of the three independent parameters, m, k and f_s, are returned by the fit.

Notice that as the age t of the cohort increases, the prevalence function $P(t)$ in Equation (8) *saturates* at the value of $f_s \leq 1$. Thus, the model prevalence function automatically satisfies the physical requirement that its value never exceed one. Thus, *both* the ordered and independent mutation models are *naturally saturated*, as they must be to be physically realistic.

3. AD incidence data

The most extensive analysis of the incidence of AD comes from the Rochester, Minnesota study [1,2]. The original data was compiled over a 25-year period (1960–1984) by Kokmen *et al.* [1]. This data was reanalysed and extended by Rocca *et al.* [2] for the last 10 years of the study (1975–1984). The modelling of AD in this paper will be based on the data in the Rocca [2] reanalysis.

The incidence of AD in men and women for Rochester, Minnesota in Ref. [2] is reproduced here in Tables 1(a) and 2(a), respectively, for convenience. The fractions that appear in these tables are incidences over 5-year intervals. To get *annual* incidences, which are required in the modelling in this paper, these fractions must be divided by 5. Since females live longer than males, the female cohorts in Table 2(a) are substantially larger than the corresponding male cohorts appearing in Table 1(a), especially above 80 years of age; thus, the female AD data is more reliable than the male data. The incidence data in these tables differ from the incidence data in Ref. [1] in one important respect: the population numbers appearing in the denominators in Tables 1(a) and 2(a) were those within an age interval that *were initially free of dementia*. In the incidence data in Ref. [1], *all* residents of Rochester, Minnesota within an age interval were included in the denominators of the AD incidence tables, *including those that were previously diagnosed with dementia*. Since the goal of the modelling in this paper is to compute the prevalence of AD as a function of age, it will be necessary to convert the form of incidence data in Ref. [2] back into the form for the incidence data used in Ref. [1].

If H_T is a random sample of a risk population at birth (age $t = 0$) and $H_{AD}(t)$ is the number within this population that has developed AD at age t, then the prevalence of AD is defined as $P(t) = H_{AD}(t)/H_T$. The incidence rate of AD at any age t used in Ref. [1] was defined as $(\mathrm{d}H_{AD}(t)/\mathrm{d}t)/H_T = \mathrm{d}P(t)/\mathrm{d}t$, which will be denoted by $\mathrm{IR}(t)$ for short. If this expression is applied to any age interval for AD, then the denominator H_T can be approximated by the total number of city residents with ages falling within this interval and the numerator is the annual number of these residents who developed AD with ages within this interval. Although not all people diagnosed with dementia have AD, in the case of Rochester, Minnesota, the great majority did. Thus, to a very good approximation of the data, it will be assumed that all dementia cases are cases of AD. The error induced in the modelling results by making this simplifying assumption is certainly far less than the observed variation in the incidence of AD from one year to the next.

Table 1(a). Male AD incidence data per 5-year age interval from Rocca *et al.* [2].

	Male AD incidence* IR*(t) over 5-year age intervals[a]							
Year	60–64 years $\theta_{0,1}$	65–69 years $\theta_{1,2}$	70–74 years $\theta_{2,3}$	75–79 years $\theta_{3,4}$	80–84 years $\theta_{4,5}$	85–89 years $\theta_{5,6}$	90–94 years $\theta_{6,7}$	95–99 years $\theta_{7,8}$
1975	1/838*	1/604	4/477	2/363	1/199	2/93	1/32	0/6
1976	0/848	0/611	2/481	2/366	7/201	5/93	2/32	0/6
1977	2/859	3/622	2/486	2/368	4/201	2/97	0/31	0/6
1978	0/873	1/630	3/488	1/374	1/204	3/100	0/29	0/6
1979	0/885	0/640	1/494	2/374	1/204	1/104	4/29	0/6
1980	0/895	1/649	2/498	1/380	3/206	4/105	2/29	1/6
1981	1/906	3/676	2/516	4/390	5/215	5/111	0/32	0/6
1982	1/918	1/700	2/536	3/400	6/221	3/111	4/33	0/7
1983	0/928	0/728	1/553	6/406	2/227	1/118	0/32	0/8
1984	1/939	2/751	1/572	4/416	4/233	3/120	0/33	0/9
Average	6.77×10^{-4}	1.81×10^{-3}	4.00×10^{-3}	6.92×10^{-3}	1.60×10^{-2}	2.79×10^{-2}	4.21×10^{-2}	1.66×10^{-2}
Standard Dev.	8.01×10^{-4}	1.72×10^{-3}	2.04×10^{-3}	3.73×10^{-3}	1.03×10^{-2}	1.44×10^{-2}	5.32×10^{-2}	5.27×10^{-2}

*The denominators in this table are the populations within an age interval that were initially free of dementia.
[a] The *annual* incidence rates within each 5-year (y) time interval is $\theta_{i,j}/5$.

In contrast to the definition of the incidence rate used in Ref. [1], the incidence rate used in Ref. [2] replaced H_T, which includes previously diagnosed cases of AD, with $H^*_{AD}(t) \equiv H_T - H_{AD}(t)$ which does not. Thus, the incidence rate in Ref. [2] was defined as $(dH_{AD}(t)/dt/H^*_{AD}(t)$, which will be denoted as IR$*(t)$ for short. The connection between these two different incidence rates is given by

$$IR(t) = IR^*(t)\frac{H^*_{AD}(t)}{H_T} = IR^*(t)\frac{[H_T - H_{AD}(t)]}{H_T} = IR^*(t)[1 - P(t)] = \frac{dP(t)}{dt}. \tag{10}$$

The last Equation in (10) can be rewritten as

$$IR^*(t) = -\frac{d}{dt}\ln[1 - P(t)]. \tag{11a}$$

Since the data shows that $P(t) = 0$ for $t \leq 60$ years, Equation (11a) can be integrated to give

$$P(t) = 1 - \exp[-\theta(t)], \quad \text{where } \theta(t) \equiv \int_{60\,y}^{t} IR^*(t)dt. \tag{11b}$$

The following relationship between the two incidence rates then follows from (10) and (11b):

$$IR(t) \equiv \frac{dP(t)}{dt} = IR^*(t)\exp[-\theta(t)]. \tag{12}$$

Since the dimensionless quantity $\theta(t) \geq 0$, it is always true that $IR(t) \leq IR^*(t)$. Using the data in Tables 1(a) and 2(a), the male and female AD incidence rates IR(t) and prevalences $P(t)$ can be computed from the incidence rates IR$^*(t)$ appearing in Ref. [2].

The male prevalences $P(t)$ are shown in Table 1(b1),(b2), while the female prevalences are shown in Table 2(b1),(b2). The prevalences appearing in these tables were calculated using Equation (11b) where the time-dependent parameter $\theta(t)$ appearing in (11b) is computed from the fractions $\theta_{i,j}$ in these tables as follows: $\theta(65) = \theta_{0,1}$, $\theta(70) = \theta_{0,1} + \theta_{1,2}$, $\theta(75) = \theta_{0,1} + \theta_{1,2} + \theta_{2,3}$, etc.

The male *annual* incidence rates IR(t) are shown in Table 1(c) while the female *annual* incidence rates are shown in Table 2(c). The *annual* incidence rates appearing in these tables were computed using Equation (12) where for example $IR^*(62.5) = \theta_{0,1}/5$ and $\theta(62.5) = \theta_{0,1}/2$, so that $IR(62.5) = \theta_{0,1}5^{-1}\exp[-\theta_{0,1}/2]$. Similarly, $IR^*(67.5) = \theta_{1,2}/5$ and $\theta(67.5) = \theta_{0,1} + \theta_{1,2}/2$, so that $IR(67.5) = \theta_{1,2}5^{-1}\exp[-\theta_{0,1} - \theta_{1,2}/2]$. Continuing in this way generates all of the annual incidence rates appearing in Tables 1(c) and 2(c). Multiplying the *annual* incidence rates in Tables 1(c) and 2(c) by 5 gives the 5-year interval incidence rates IR(t) that can be compared with the 5-year interval incidences rates IR$^*(t)$ given in Tables 1(a) and 2(a); indeed, the ratio of 5-year interval incidence rates IR(t)/IR$^*(t)$ computed from these tables are in agreement with the IR(t) and IR$^*(t)$ curves appearing in Figure 1 in Ref. [2].

In the language of the models constructed here, remembering that m mutations are required to initiate AD, the prevalence $P(t)$ of AD at age t is given by the probability that a brain neuron has died (see (5), (7), or (A6) in the Appendix).

So, what do that annual AD incidence IR(t) and prevalence $P(t)$ data appearing in tables say about the disease?

Table 1(b1). Male AD prevalence as a function of age (t) in years computed directly from the Rocca data in Table 1(a) using Equation (11b).

	Male AD prevalence $P(t)$ at age t (in years) computed from Rocca data in Table 1(a)							
Year	$P(65)$	$P(70)$	$P(75)$	$P(80)$	$P(85)$	$P(90)$	$P(95)$	$P(100)$
1975	1.192×10^{-3}	0.002844	0.01117	0.01660	0.02153	0.04235	0.07181	0.07181
1976	0	0	0.004149	0.009576	0.04347	0.09354	0.1484	0.1484
1977	0.002325	0.007125	0.01120	0.01656	0.03594	0.05561	0.05561	0.05561
1978	0	0.001586	0.007705	0.01035	0.01519	0.04429	0.04429	0.04429
1979	0	0	0.002022	0.007344	0.01219	0.02165	0.1477	0.1477
1980	0	0.001539	0.005541	0.008155	0.02249	0.05903	0.1217	0.2565
1981	0.001103	0.005526	0.009393	0.01948	0.04202	0.08421	0.08421	0.08421
1982	0.001088	0.002514	0.006229	0.01365	0.04007	0.06567	0.1723	0.1723
1983	0	0	0.001806	0.01645	0.02507	0.03330	0.03330	0.03330
1984	0.001064	0.003721	0.005461	0.01497	0.03174	0.05565	0.05565	0.05565
Average	0.0006774	0.002485	0.006466	0.01331	0.02897	0.05553	0.09351	0.1070
Standard Deviation	0.0008011	0.002419	0.003388	0.004178	0.01127	0.02192	0.04993	0.07181

Table 1(b2). Male AD prevalence as a function of age (t) in years computed directly from the Rocca data in Table 1(a) using Equation (11b).

	Male AD prevalence $P(t)$ at age t (in years) computed from Rocca data in Table 1(a)							
Year	$P(62.5)$	$P(67.5)$	$P(72.5)$	$P(77.5)$	$P(82.5)$	$P(87.5)$	$P(92.5)$	$P(97.5)$
1975	0.0005964	0.002019	0.007017	0.01389	0.01907	0.03199	0.05719	0.07181
1976	0	0	0.002076	0.006866	0.02667	0.06884	0.1214	0.1484
1977	0.001163	0.004728	0.009166	0.01388	0.02630	0.04582	0.05561	0.05561
1978	0	0.0007933	0.004650	0.009030	0.01277	0.02985	0.04429	0.04429
1979	0	0	0.001011	0.004687	0.009774	0.01693	0.08685	0.1477
1980	0	0.0007701	0.003542	0.006849	0.01535	0.04093	0.09092	0.1919
1981	0.0005517	0.0033172	0.007451	0.01444	0.03081	0.06335	0.08421	0.08421
1982	0.0005445	0.001802	0.004374	0.009949	0.02695	0.05295	0.1206	0.1723
1983	0	0	0.0009037	0.009155	0.02077	0.02920	0.03330	0.03330
1984	0.0005323	0.002393	0.004591	0.01023	0.02339	0.04377	0.05565	0.05565
Average	0.0003388	0.001582	0.004478	0.009898	0.02118	0.04236	0.07501	0.1005
Standard Deviation	0.0004007	0.001584	0.002764	0.003325	0.006886	0.01617	0.03066	0.05853

Table 1(c). Male annual AD incidence rate IR(t) computed from IR*(t) data in Table 1(a) using Equation (12).

	Male *annual* AD incidence rate IR(t) at indicated age (t) in years							
Year/age	62.5	67.5	72.5	77.5	82.5	87.5	92.5	97.5
1975	2.385×10^{-4}	3.304×10^{-4}	1.665×10^{-3}	1.086×10^{-3}	9.858×10^{-4}	4.163×10^{-3}	5.892×10^{-3}	0
1976	0	0	8.298×10^{-4}	1.085×10^{-3}	6.779×10^{-3}	1.001×10^{-2}	1.098×10^{-2}	0
1977	4.651×10^{-4}	9.600×10^{-4}	8.155×10^{-4}	1.071×10^{-3}	3.875×10^{-3}	3.934×10^{-3}	0	0
1978	0	3.172×10^{-4}	1.223×10^{-3}	5.299×10^{-4}	9.678×10^{-4}	5.820×10^{-3}	0	0
1979	0	0	4.044×10^{-4}	1.064×10^{-3}	9.708×10^{-4}	1.890×10^{-3}	2.519×10^{-2}	0
1980	0	3.079×10^{-4}	8.003×10^{-4}	5.227×10^{-4}	2.867×10^{-3}	7.307×10^{-3}	1.253×10^{-2}	2.693×10^{-2}
1981	2.206×10^{-4}	8.846×10^{-4}	7.694×10^{-4}	2.021×10^{-3}	4.507×10^{-3}	8.438×10^{-3}	0	0
1982	2.177×10^{-4}	2.852×10^{-4}	7.430×10^{-4}	1.485×10^{-3}	5.283×10^{-3}	5.119×10^{-3}	2.131×10^{-2}	0
1983	0	0	3.613×10^{-4}	2.928×10^{-3}	1.725×10^{-3}	1.645×10^{-3}	0	0
1984	2.128×10^{-4}	5.313×10^{-4}	3.480×10^{-4}	1.903×10^{-3}	3.353×10^{-3}	4.781×10^{-3}	0	0
Average	1.354×10^{-4}	3.616×10^{-4}	7.961×10^{-4}	1.370×10^{-3}	3.131×10^{-3}	5.311×10^{-3}	7.592×10^{-3}	2.693×10^{-3}
Stand. Dev.	1.602×10^{-4}	3.439×10^{-4}	4.064×10^{-4}	7.392×10^{-4}	2.012×10^{-3}	2.683×10^{-3}	9.575×10^{-3}	8.517×10^{-3}

Table 2(a). Female AD incidence data per 5-year age interval from Rocca *et al.* [2].

Year	Female AD incidence* IR*(t) over 5-year age interval [a]							
	60–64 years $\theta_{0,1}$	65–69 years $\theta_{1,2}$	70–74 years $\theta_{2,3}$	75–79 years $\theta_{3,4}$	80–84 years $\theta_{4,5}$	85–89 years $\theta_{5,6}$	90–94 years $\theta_{6,7}$	95–99 years $\theta_{7,8}$
1975	0/1086*	0/999	3/923	10/735	12/509	8/284	2/96	1/24
1976	0/1087	0/1005	5/936	7/751	15/524	10/290	2/101	0/24
1977	0/1089	1/1013	6/950	11/769	10/539	12/296	2/107	0/25
1978	1/1088	0/1019	1/968	2/785	9/554	14/302	5/112	1/26
1979	0/1090	1/1025	3/983	2/804	7/569	7/308	8/118	0/28
1980	1/1091	3/1033	4/997	11/821	16/584	15/314	6/124	0/28
1981	0/1107	1/1044	4/1001	8/834	14/599	8/326	3/131	2/28
1982	0/1119	0/1056	8/1005	6/844	11/611	12/339	4/141	1/28
1983	1/1134	0/1068	1/1009	5/859	14/623	10/349	4/147	4/28
1984	0/1146	1/1078	0/1013	4/870	16/637	13/362	6/157	3/28
Average	2.71×10^{-4}	6.75×10^{-4}	3.59×10^{-3}	8.27×10^{-3}	2.15×10^{-2}	3.44×10^{-2}	3.36×10^{-2}	4.37×10^{-2}
Standard Deviation	4.37×10^{-4}	9.17×10^{-4}	2.50×10^{-3}	4.48×10^{-3}	5.16×10^{-3}	8.61×10^{-3}	1.59×10^{-2}	4.98×10^{-2}

*The denominators in this table are the populations within an age interval that were initially free of dementia.
[a] The *annual* incidence rates within each 5-year (y) time interval is $\theta_{i,j}/5$.

Table 2(b1). Female AD prevalence as a function of age (t) in years computed directly from the Rocca data in Table 2(a) using Equation (11b).

	Female AD prevalence $P(t)$ at age t (in years) computed from Rocca data in Table 2(a)							
Year	$P(65)$	$P(70)$	$P(75)$	$P(80)$	$P(85)$	$P(90)$	$P(95)$	$P(100)$
1975	0	0	3.245×10^{-3}	0.01671	0.03962	0.06630	0.08555	0.1228
1976	0	0	5.327×10^{-3}	0.01455	0.04236	0.07482	0.09296	0.09296
1977	0	9.866×10^{-4}	7.276×10^{-3}	0.02137	0.03936	0.07753	0.09461	0.09461
1978	9.187×10^{-4}	9.187×10^{-4}	1.950×10^{-3}	0.004489	0.02053	0.06490	0.1057	0.1394
1979	0	9.751×10^{-4}	4.019×10^{-3}	0.006493	0.01864	0.04069	0.1035	0.1035
1980	9.161×10^{-4}	3.813×10^{-3}	7.802×10^{-3}	0.02100	0.04746	0.09189	0.1347	0.1347
1981	0	9.574×10^{-4}	4.941×10^{-3}	0.01444	0.03720	0.06054	0.08181	0.1451
1982	0	0	7.928×10^{-3}	0.01495	0.03253	0.06617	0.09229	0.1241
1983	8.814×10^{-4}	8.814×10^{-4}	1.871×10^{-3}	0.007664	0.02971	0.05712	0.08243	0.2045
1984	0	9.272×10^{-4}	9.272×10^{-4}	0.005510	0.03017	0.06438	0.09947	0.1909
Average	2.716×10^{-4}	9.460×10^{-4}	4.528×10^{-3}	0.01272	0.03376	0.06643	0.09732	0.1353
Stand. Dev.	4.374×10^{-4}	1.101×10^{-3}	2.565×10^{-3}	0.006279	0.009272	0.01346	0.01553	0.03763

Table 2(b2). Female AD prevalence as a function of age (t) in years computed directly from the Rocca data in Table 2(a) using Equation (11b).

	Female AD prevalence $P(t)$ at age t (in years) computed from Rocca data in Table 2(a)							
Year	$P(62.5)$	$P(67.5)$	$P(72.5)$	$P(77.5)$	$P(82.5)$	$P(87.5)$	$P(92.5)$	$P(97.5)$
1975	0	0	0.001623	0.01000	0.02823	0.05305	0.07597	0.1044
1976	0	0	0.002667	0.009952	0.02856	0.05873	0.08393	0.09296
1977	0	0.0004934	0.004136	0.01435	0.03041	0.05864	0.08611	0.09461
1978	0.0004594	0.0009187	0.001434	0.003220	0.01254	0.04297	0.08554	0.1227
1979	0	0.0004876	0.002498	0.005257	0.01258	0.02973	0.07266	0.1035
1980	0.0004581	0.002365	0.005809	0.01442	0.03432	0.06994	0.1136	0.1347
1981	0	0.0004788	0.002951	0.009706	0.02589	0.04895	0.07124	0.1140
1982	0	0	0.003972	0.01144	0.02378	0.04950	0.07933	0.1083
1983	0.0004408	0.0008814	0.001376	0.004771	0.01875	0.04351	0.06986	0.1456
1984	0	0.0004637	0.0009272	0.003221	0.01792	0.04743	0.08209	0.1464
Average	0.0001358	0.0006089	0.002739	0.008635	0.02330	0.05024	0.08203	0.1167
Stand. Dev.	0.0002187	0.0007020	0.001531	0.004266	0.007537	0.01089	0.01258	0.01983

Table 2(c). Female annual AD incidence rate IR(t) computed from IR*(t) data in Table 2(a) using Equation (12).

	Female *annual* AD incidence rate IR(t) at indicated age (t) in years							
Year/age	62.5	67.5	72.5	77.5	82.5	87.5	92.5	97.5
1975	0	0	6.490×10^{-4}	2.693×10^{-3}	0.004582	0.005334	0.003850	0.007463
1976	0	0	1.065×10^{-3}	1.845×10^{-3}	0.005561	0.006491	0.003628	0
1977	0	1.973×10^{-4}	1.257×10^{-3}	2.819×10^{-3}	0.003597	0.007632	0.003416	0
1978	1.837×10^{-4}	0	2.063×10^{-4}	5.079×10^{-4}	0.003208	0.008873	0.008164	0.06748
1979	0	1.950×10^{-4}	6.088×10^{-4}	4.949×10^{-4}	0.002429	0.004410	0.01257	0
1980	1.832×10^{-4}	5.794×10^{-4}	7.977×10^{-4}	2.641×10^{-3}	0.005291	0.008885	0.008578	0
1981	0	1.914×10^{-4}	7.968×10^{-4}	1.899×10^{-3}	0.004553	0.004667	0.004253	0.01265
1982	0	0	1.585×10^{-3}	1.405×10^{-3}	0.003515	0.006729	0.005223	0.006368
1983	1.762×10^{-4}	0	1.979×10^{-4}	1.158×10^{-3}	0.004410	0.005481	0.005062	0.02440
1984	0	1.854×10^{-4}	0	9.165×10^{-4}	0.004933	0.006841	0.007015	0.01829
Average	5.432×10^{-5}	1.348×10^{-4}	7.165×10^{-4}	1.638×10^{-3}	0.004208	0.006534	0.006176	0.007593
Stand. Dev.	8.749×10^{-5}	1.831×10^{-4}	4.991×10^{-4}	8.830×10^{-4}	9.915×10^{-4}	1.594×10^{-3}	0.002919	0.008535

Table 3. Age-specific AD incidence rates (% per year) *in cohorts initially free of dementia*.

Age range (in years)	Rochester, Minnesota study [2]		Baltimore Longitudinal study [3]		East Boston, Mass. study [4]
	Men	Women	Men	Women	
55–59	0	0	0	0	
60–64	0.067	0.027	0	0.25	
65–69	0.181	0.067	0.09	0.22	0.6
70–74	0.400	0.359	0.55	0.17	1.0
75–79	0.692	0.827	0.75	1.10	2.0
80–84	1.60	2.5	1.25	3.61	3.3
85–89/(85+)	2.79/(3.00)	3.44/(3.49)	(7.2)	(5.27)	(8.4)
90–94	4.21	3.36			
95–99	1.66	4.37			
Study population size/(year)	4,680/(1975) to 5,484/(1984)	7,140/(1975) to 7,933/(1984)	802	434	2,313

Figure 1(a) is a plot of the mean male *annual* incidence rate data IR(t) appearing in Table 1(c) together with the standard deviation error of each data point. Figure 1(b) is a similar plot of the male AD prevalence $P(t)$ data appearing in Table 1(b1),(b2). Figure 1(a),(b) clearly shows that the errors in the male data above 90 years old are so great that it is impossible to extrapolate the IR(t) and $P(t)$ data into the region above 100 years old with any confidence of accuracy. A very similar situation results from the female AD data.

Figure 2(a) is a plot of the mean female *annual* incidence rate data IR(t) appearing in Table 2(c) together with the standard deviation error of each data point. Figure 2(b) is a similar plot of the female AD prevalence $P(t)$ data appearing in Table 2(b1),(b2). Figure 2(a),(b) again shows that the errors in the data above 90 years old are so great that it is impossible to accurately extrapolate the IR(t) and $P(t)$ data into the region above 100 years old.

This is a situation where modelling the disease may help answer the question of how the AD incidence rate IR(t) and prevalence $P(t)$ functions behave in the region above 100 years of age.

4. Other studies of age-related AD incidence rates

The results of other studies of age-related AD incidence rates is shown in Table 3 along with the data for the Rochester, Minnesota study that was used for the modelling in this paper. The Rochester study was clearly more detailed than the others and benefits from having a much greater cohort population. The cohort population in the Rochester Study was about ten times greater than that of the Baltimore Longitudinal Study. Moreover, by choosing an entire city for the study, the Rochester results avoid errors stemming from choosing a sample cohort that is not representative of the population as a whole. For example, the ratio of women to men in the Rochester study is about 1.5, a result that is in general agreement with the elderly population in the USA. However, the ratio of women to men in the Baltimore Longitudinal Study is 0.54, a ratio that is not representative of the USA population as a whole. Finally, the Rochester Study is the only one of the three studies shown in Table 3 that has detailed data for the 85–89, 90–94 and 95–99 year age

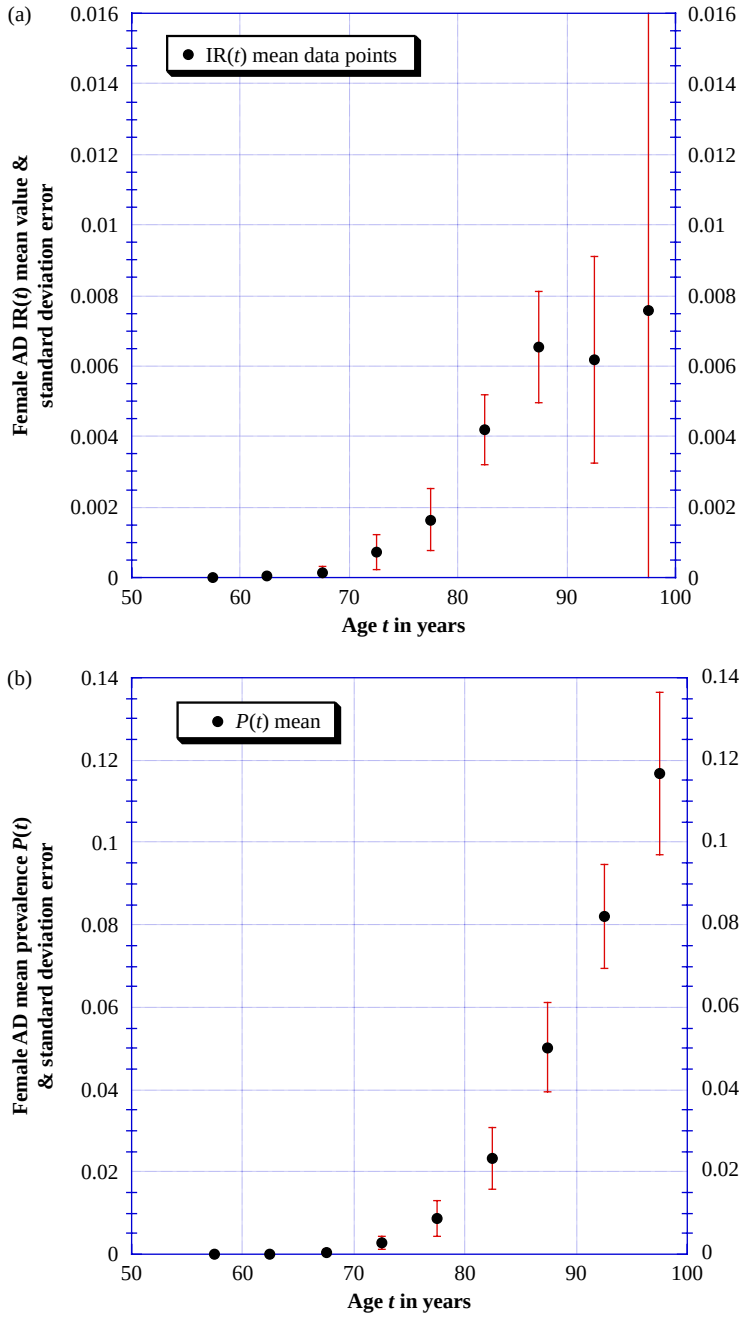


Figure 2. (a) Female AD data for annual incidence rate IR(t) mean value and standard deviation error as a function of t in years from Table 2 (c) and (b) Female AD data for mean prevalence $P(t)$ and standard deviation error as a function of age t in years from Table 2 (b2).

groups. For all of these reasons, the data of the Rochester Study was deemed more reliable and was used in the modelling in this paper. Nonetheless, the corresponding AD incidence rate data between any two studies in Table 3 are within a factor of 2 of each other in general. Thus, the general features of the model results using any of these three data sets are very similar and differ only in how fast the AD prevalence rate rises with age.

5. Modelling the AD incidence rate curve

The AD incidence rate curve produced by the ordered mutation model is given by Equation (A8) in the Appendix. Setting the susceptibility fraction parameter f_s equal to 1 and fitting this function to the male AD data in Table 1(c) yields the fit shown in Figure 3(a). To achieve this fit, the mean data point at 97.5 years was ignored since this point was deemed too unreliable. Nonetheless, the model incidence rate function IR(t) returned by the fit falls within one standard deviation of not only the 97.5 year data point but of *all of the data points* as shown in Figure 3(a). In this fit, the entire male population was assumed to be vulnerable to developing AD ($f_s = 1$). The values of k and m, the only remaining independent parameters, were determined by the fit and are shown in Figure 3(a). The projection of the IR(t) model function into the region above 100 years old is also shown in Figure 3(a). Notice that the IR(t) curve peaks at around 120 years old and monotonically declines thereafter towards zero. The projected values of the model incidence rate function IR(t) in the region $t < 60$ years old are vanishingly small, in agreement with experiment; indeed, this is one of the strongest features of *both* the ordered and independent mutation models.

Letting f_s become an independent parameter and repeating the ordered mutation model fit gives the result shown in Figure 3(b). The fit to the data here is better than the one in Figure 3(a), with the (chisq) error [see (A11) in the Appendix] reduced from what it is in Figure 3(a) by 35.5%. As before, the model incidence rate curve falls within one standard deviation of all of the data points. However, this fit predicts that only 45.9% of men are susceptible to acquiring AD ($f_s = 0.459$), and, therefore, that 54.1% are *immune* to getting it. This fit also predicts that the average number of mutations necessary for a brain cell associated with AD to spontaneously die is $m = 45$, and the average elapsed time between consecutive mutations is $T = k^{-1} = (0.414)^{-1}\ y = 2.41$ years. The incidence rate function IR(t) extrapolated into the non-fitted region peaks around 105 years and monotonically declines towards zero thereafter.

Looking carefully at the fit in Figure 3(b) shows that it is particularly poor for the early incidence data points. To improve the fit to this data, a more sophisticated ordered model will be constructed where it will be assumed that, not one, but two susceptible groups are responsible for generating all of the AD data. It will be assumed that, because of genetic inheritance, a fraction of the population is born with a head-start in development of AD in that it is born with many of the mutations needed to cause it. Thus, it will be assumed that the male AD incidence data is generated by a compound incidence function given by

$$I(t) = I_1(f_{s1}, k, m_1) + I_2(f_{s2}, k, m_2), \tag{13}$$

where the notation in Equation (A8) in the Appendix has been used. Thus, the compound model assumes that there are two different groups of the population that are susceptible to acquiring AD, one that has model parameters (f_{s1}, k and m_1) and the second that has model parameters (f_{s2}, k and m_2). Notice that it is assumed that the mutation rates k for both susceptible groups are identical, and it will be assumed that $m_1 < m_2$ so that

$I_1(f_{s1}, k \text{ and } m_1)$ describes the group born with some common average number (not zero) of AD mutations. Fitting the male AD incidence data with the compound incidence function in Equation (13) gives the much-improved result shown in Figure 3(c), with a fit error (chisq) that is 60.8% lower than the error in Figure 3(b). Interestingly, the fit returns the values $f_{s1} = 0.01163$ and $f_{s2} = 0.2465$ so that $f_{s1} = 0.0472 f_{s2}$. Thus, a particular small fraction of the population (about 1%) is predicted to be born with AD mutations on the average, certainly a plausible result. The fit also produces the values $m_1 = 62$ and $m_2 = 79$ so that the particularly susceptible population is born with an average of $79 - 62 = 17$ AD mutations. The projection of the model incidence function returned by this fit is also shown in Figure 3(c) along with the standard deviation error in the data points.

Integrating the model incidence function shown in Figure 3(c) gives the model prevalence function which is given as a series of points in the second column in Table 4. The prevalence curve must be a monotonically increasing function of age t, and it musty saturate at the value of $f_{s1} + f_{s2} = 0.258136$, which, as seen in Table 4, it certainly does. Thus, the results in Table 4 constitute a check on the modelling results in Figure 3(c).

A measure of the error of the least-squares fit is given by the sum (chisq) given in (A11) of the Appendix. A completely different positive definite sum that measures the statistical quality of a fit is known as the chi-square distribution goodness-of-fit test. Even though these two tests have virtually the same name, they are not to be confused with each other. The last part of the Appendix contains a description of the chi-square distribution goodness-of-fit test and all of the details of the application of this test to the fit in Figure 3(c).

Since there are eight data points in the fit in Figure 3(c), there are $8-1=7$ degrees of freedom in this problem. The positive definite sum in the chi-square distribution goodness-of-fit test will be denoted by χ^2, and for the fit in Figure 3(c), it is found that $\chi^2 = 3.83$. For a problem with 7 degrees of freedom, a χ^2 value of $\chi^2 = 3.83$ means that there is a 79.8% probability (the p-value is 0.798) that the expected (model) and observed (data) distributions are statistically identical. Thus, by any conventionally used criterion, the model fit in Figure 3(c) is very good.

It is interesting to see what sort of fit to the same data is returned by the independent mutation model. Using the incidence rate function given in Equation (9) gives the fit shown in Figure 4(a). Here, choosing the susceptible fraction parameter to equal $f_s = 1$ gives a fit with the *least* error. Thus, no immunity to AD in the male population is predicted by the independent mutation model. Although the error of the fit here is about 10% higher than the fit in Figure 3(b) for the ordered mutation model, this model is nonetheless about as credible. Notice that the projected values of the independent model incidence rate function IR(t) in Figure 4(a) in the region $t < 60$ years old are vanishingly small in agreement with experiment. As has already been pointed out, this is one of the strongest features of *both* the ordered and independent mutation models.

To improve the fit of the independent mutation model, especially for the early points, a compound version of this model will be tried in exactly the fashion outlined in Equation (13) for the ordered model. The compound independent model fit is shown in Figure 4(b), and it has reduced the fit error (chisq) by 75.5% from that in Figure 4(a). The most important result of this fit is that this model now predicts widespread immunity to AD. Since the fit returns the values $f_{s1} = 0.0359$ and $f_{s2} = 0.329$, the compound independent model predicts that 63.4% of men are immune to acquiring AD.

Table 4. Prevalence curve computed from the compound ordered model fitted to AD male and female data.

	$P(t)$ curve for AD computed from compound ordered model fits	
Age t (in years)	Male	Female
60	0.00020017	0.00012724
62.5	0.00040647	0.00022508
65	0.00075842	0.00038961
67.5	0.00131803	0.00068675
70	0.00216453	0.00125721
72.5	0.00340978	0.00236271
75	0.00522482	0.00442234
77.5	0.00787097	0.00801061
80	0.01171971	0.01379102
82.5	0.01724228	0.02237979
85	0.02495481	0.03416782
87.5	0.03532024	0.04915843
90	0.04862735	0.06688494
92.5	0.06488146	0.08644810
95	0.08374332	0.10666889
97.5	0.10453950	0.12630920
100	0.12634750	0.14429146
102.5	0.14813128	0.15985641
105	0.16889489	0.17262767
110	0.20432751	0.18998329
115	0.22928373	0.19878664
120	0.24435096	0.20254407
125	0.25224469	0.20392637
130	0.25587292	0.20434332
135	0.25735047	0.20446173
140	0.25788830	0.20449045
145	0.25806467	0.20449664
150	0.25811714	0.20449785

Because the fit errors of the compound ordered and independent models are so close to each other (compare Figures 3(c) and 4(b)), it is impossible to decide which of these two models is more credible. However, both models predict widespread immunity to AD in the male population.

Fitting the female AD incidence data by the ordered mutation model with the value of the susceptible fraction set equal to $f_s = 1$ gives the results plotted in Figure 5(a). Notice that the mean value of the data point at $t = 92.5$ years was left off of the fit since it was deemed improbable – the mean incidence rate of this point was *lower* than that of the rate on either side of it. The fit here is determined by only two independent parameters (m and k) whose values are shown in Figure 5(a), and the (chisq) error of the fit is relatively large. Notice that the fit to the earlier data points is particularly poor, lying outside one standard deviation from the mean values of these points.

Repeating the above fit with f_s now being an independent parameter leads to the convincing fit shown in Figure 5(b). The (chisq) error of the fit in Figure 5(b) with $f_s = 0.205$ is over 62 times *smaller* than the fit in Figure 5(a) with $f_s = 1$. Again, the data point at $t = 92.5$ years was left off of the fit, but nonetheless the resulting model incidence rate function IR(t) lies within one standard deviation of the mean for *all* of the data points,

Table 5. Total number of AD cases in the USA in 2000.

Age range (in years)	Male population[a]	Male prevalence $P(t)$ from Table 1(b2)	Male AD cases	Female population[a]	Female prevalence $P(t)$ from Table 2(b2)	Female AD cases
60–64	5,165,703	0.00033885	1,750	5,699,027	0.00013585	774
65–69	4,402,844	0.0015824	6,967	5,131,111	0.00060897	3,124
70–74	3,904,321	0.0044786	17,485	4,945,625	0.0027398	13,550
75–79	3,051,227	0.0098989	30,203	4,374,151	0.0086356	37,773
80–84	1,854,596	0.021189	39,297	1,754,838	0.023301	40,889
85–89	1,099,019	0.042369	46,564	1,690,798	0.050249	84,960
90–94	438,269	0.075011	32,874	674,261	0.082038	55,315
95–99	112,975	0.10054	11,358	173,808	0.11676	20,293
Over 100	19,875	0.14813 [b]	2,944	30,578	0.15985[b]	4,888
Gender AD sum			189,442			261,566
AD total				451,008		

[a] In year 2000 from US Census Bureau.
[b] Computed prevalence at age 102.5 years from Table 4 [Total USA population in 2000 was 282,338,631].

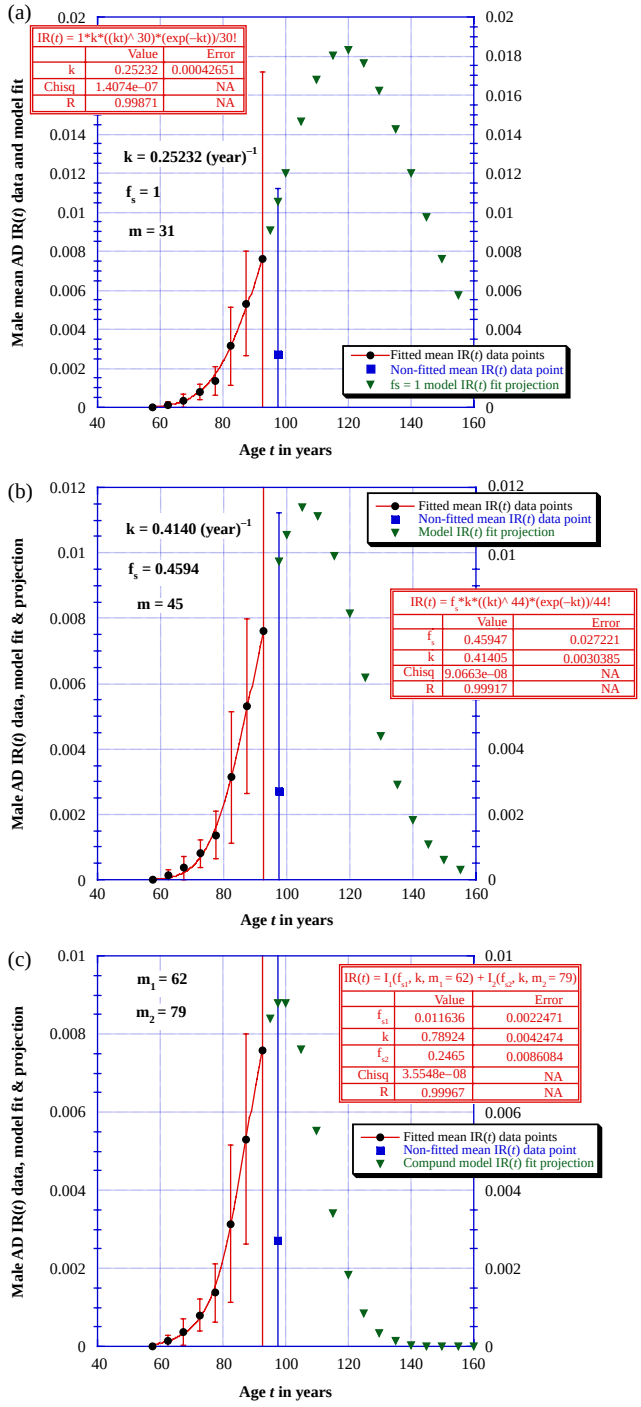


Figure 3. (a) Male AD mean incidence rate IR(t) data and fit using $f_s = 1$ ordered mutation model. (b) Male AD incidence rate IR(t) data (mean and standard deviation error) as a function of age t together with ordered mutation model fit and projection. (c) Male AD incidence rate IR(t) data (mean and standard deviation error) as a function of age t together with COMPOUND ordered mutation model fit and projection.

as seen in Figure 5(b). This fit predicts that 79.4% of females are *immune* to developing AD, the number of mutations necessary for a brain cell associated with AD to spontaneously die is $m = 87$, and the average time between consecutive mutations is $T = k^{-1} = [0.916]^{-1}y = 1.09$ years. Figure 5(b) also shows the projection of the model

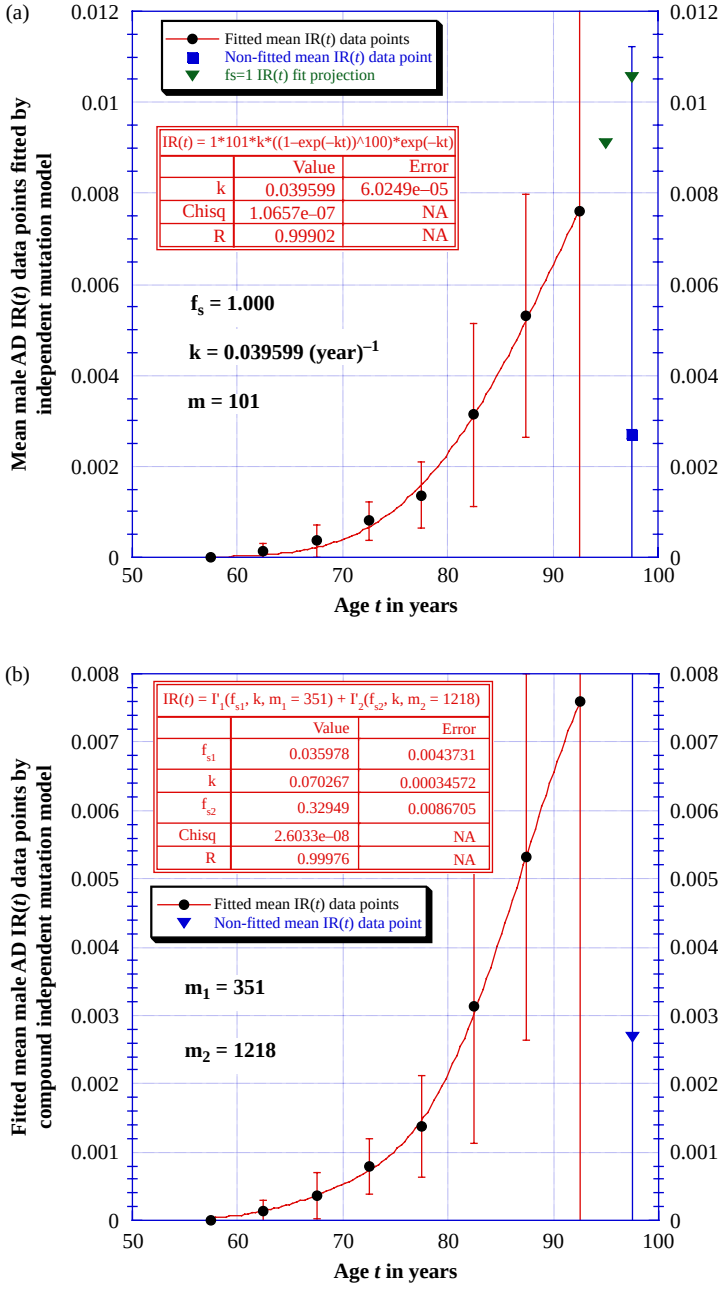


Figure 4. (a) Male AD incidence rate IR(t) data and $f_s = 1$ independent mutation model fit. (b) Compound independent mutation model fit to male AD incidence rate IR(t) data.

incidence rate IR(t) curve into the region above 100 years of age. Notice that the model incidence curve peaks around 98 years of age and monotonically declines thereafter.

To possibly improve the fit in Figure 5(b), a fit using the compound model incidence function in (13) will now be executed. The improved result is shown in Figure 5(c), with a fit error (chisq) that is 6.80% lower than the error in Figure 5(b). In the female data case, the fit returns the values $f_{s1} = 0.001078$ and $f_{s2} = 0.2034$ so that $f_{s1} = 0.0529 f_{s2}$, virtually the same result as for the male data. Thus, once again, a particular small fraction of the considered population (about 0.1%) is predicted to be born with AD mutations on the average. The fit also produces the values $m_1 = 66$ and $m_2 = 89$ so that the particularly susceptible female population is born with an average of $89 - 66 = 23$ AD mutations, slightly higher than it was for the male population. The projection of the model incidence function returned by this fit is also shown in Figure 5(c) along with the standard deviation error in the data points.

Integrating the model incidence function shown in Figure 5(c) gives the model prevalence function which is given as a series of points on the right-hand side in Table 4. The prevalence curve must be a monotonically increasing function with age t, and it must saturate at the value of $f_{s1} + f_{s2} = 0.204498$, which, as seen in Table 4, it certainly does. Thus, the results in Table 4 again constitute a check on the modelling results in Figure 5(c), and the modelling passes this test.

Applying the chi-square distribution goodness-of-fit test to the fit in Figure 5(c) gives the results found in the last part of the Appendix summarized here.

Since there are again eight data points in the fit in Figure 5(c), there are again 8−1=7 degrees of freedom in this problem. The value of test statistic for this fit turned out to be $\chi^2 = 4.64$. For a problem with 7 degrees of freedom, a χ^2 value of $\chi^2 = 4.64$ means that there is a 70.3% probability (the p-value is 0.703) that the expected (model) and observed (data) distributions are statistically identical. Thus, once again, by any conventionally used criterion, the model fit in Figure 5(c) is very good.

Proceeding in the same way, fitting the same female AD incidence data using the *independent* mutation model yields the result in Figure 6(a). Since the value of the susceptible fraction returned by this fit is $f_s = 0.271$, this model predicts that 72. 8% of females are *naturally immune* to developing AD. In the independent model result in Figure 6(a), the number of mutations necessary for a brain cell associated with AD to spontaneously die $m = 1605$, and the average time required for a mutation to occur is $T = k^{-1} = [0.780]^{-1} y = 1.28$ years.

Fitting the same female data with the compound independent model incidence rate function gives the fit shown in Figure 6(b). This more sophisticated model reduces the fit error by 59.1% over what it was in the single term fit in Figure 6(a). Here, the values of the susceptible fractions returned by the fit are $f_{s1} = 0.0177$ and $f_{s2} = 0.239$ so that this model predicts that 74.3% of females are immune to developing AD.

Interestingly, both the compound ordered and independent mutation models predict that most females are naturally immune to developing AD. The fits of both models produce comparable errors, so it is not possible to decide on the basis of the modelling which model produces superior results. Since females live longer than males, the female AD cohorts above 70 years old in the Rochester, Minnesota study were generally 2–4 times larger than the corresponding male cohorts (see Tables 1(a) and 2(a)). Thus, the female AD data is probably more reliable than the male AD data, and, therefore, the female modelling results are expected to be more reliable than the male modelling results. The different values of the fit parameters (f_s, m and k) predicted by a mutation model for male and female cohorts demonstrates that AD develops through different pathways, as is commonly true for various cancers.

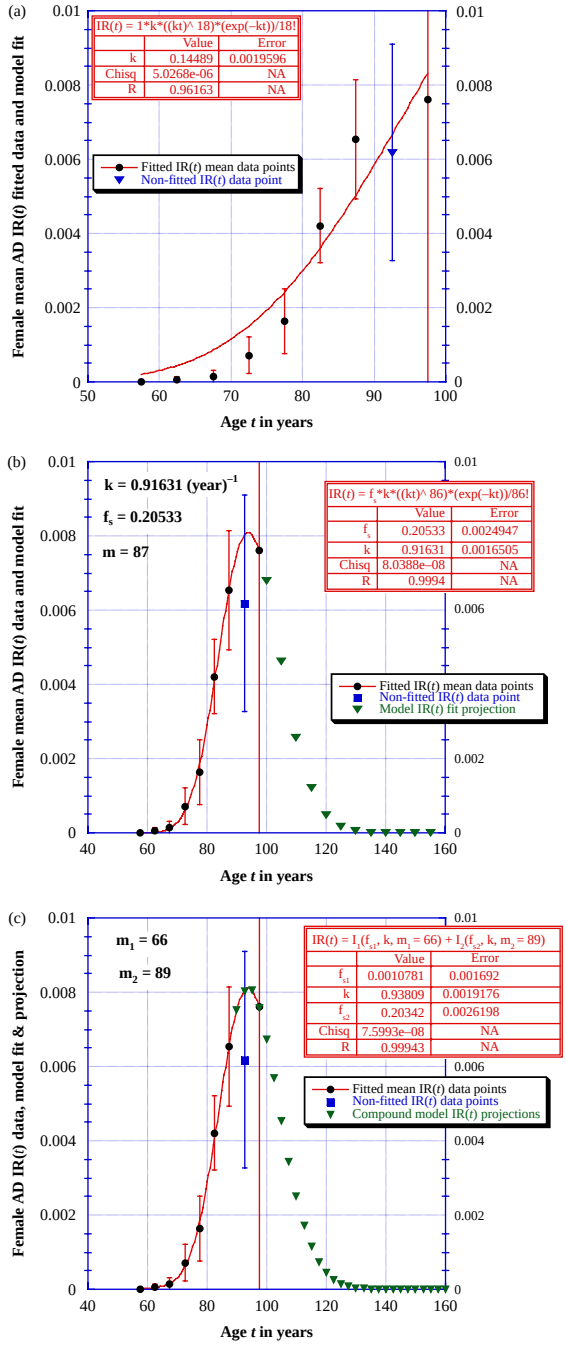


Figure 5. (a) Female AD mean incidence rate IR(t) fitted data using $f_s = 1$ ordered mutation model. (b) Female AD mean incidence rate IR(t) data and fit using ordered mutation model. (c) Female AD incidence rate IR(t) data (mean and standard deviation error) as a function of age t together with COMPOUND ordered mutation model fit and projection.

Although the causes of the mutations in the mutation models are unknown to date, viral infections are certainly a possibility.

6. How reliable is the model prediction of AD immunity?

To test the sensitivity of the value of a typical model parameter determined by the least-square fits described in this paper, the χ^2-test statistic will be calculated for the entire range of physically permitted values for the susceptible fraction parameter f_{s2} for the dominant AD-susceptible population. Again, the χ^2-test statistic is not to be confused with the least-squares fit error defined in (A11) in the Appendix which is denoted as Chisq, a very similar annotation. The dependence of *both* of these errors as a function of f_{s2} is shown in Figure 7.

The procedure followed here was to chose a particular value for f_{s2} and allow the remaining parameters of the model to be chosen by the least-squares fit to the AD data; this procedure was repeated until the entire physically permitted range of values for f_{s2} was covered. As seen in Figure 7, the minimum values for *both* errors occurs at $f_{s2} = 0.329$, where $\chi^2 = 3.12$ ($p = 0.8733$), and both errors increase in *either* direction as we move away from this value. The details of the χ^2 calculation for the point $f_{s2} = 0.329$ are shown in Table 9 in the appendix; the values of χ^2 for all the other points in Figure 7 are calculated in the same way.

The maximum physically permitted value of f_{s2} in the compound model occurs when $f_{s2} = 1 - f_{s1}$ so that the *entire* population is susceptible to acquiring AD. The value for f_{s2} in this case turns out to be $1 - 0.0111 = 0.988$ with $\chi^2 = 4.67$ ($p = 0.699$). Thus, although this value for f_{s2} is statistically less probable than the value of $f_{s2} = 0.329$ above, the difference is not great enough to rule it out. Thus, although the modelling in this paper suggests that immunity to AD *may* exist, *it does not prove it.*

At the other extreme for $f_{s2} = 0.18$ the least-squares fit for the remaining parameters produces $\chi^2 = 32.3$ ($p = 0.00004$), an extremely unlikely result. Thus, values of $f_{s2} < 0.18$ are statistically implausible, and 0.18 can be regarded as a lower bound on the value of f_{s2}.

7. Computing the number of AD patients in the USA or any other country

The age distribution of the over 60 year-old population of the USA in the year 2000 is shown in columns 2 and 5 in Table 4. These figures were obtained from the US Census Bureau. The mean prevalence of AD for each age group was taken from either Table 1(b1) or 2(b1) and is shown in either column 3 or 6 in Table 4. Multiplying columns 2 and 3 in Table 4 together yields the result in column 4 for the number of males in each age group that had AD in the USA in the year 2000; similarly, multiplying columns 5 and 6 in Table 4 together yields the result in column 7 for the number of females in each age group that had AD in the USA in the year 2000. Summing columns 4 and 7 in Table 4, the total number of people in the USA that had AD in the year 2000 is estimated to be 451,400, or about 0.16% of the population.

The prevalence results in Table 4 can be used to estimate the number of people with AD in any country or city by merely changing the numbers in columns 2 and 5 to reflect the population distribution of that country or city.

The US Census Bureau projects that in the year 2007 the number of Americans in the 65–84 year old bracket will increase by 2,328,000 and those in the 85-years-old and above bracket will increase by 1,297,000. The average age of the 65–84 year age interval is 75 years old, and from Tables 1(b1) and 2(b1) the average prevalence at age 75 years is $(1/2)[0.006466 + 0.004528] = 0.005497$. Thus, the number of AD cases in this age bracket is expected to rise by $2{,}328{,}000 \times 0.005497 = 12{,}797$ cases by the year 2007.

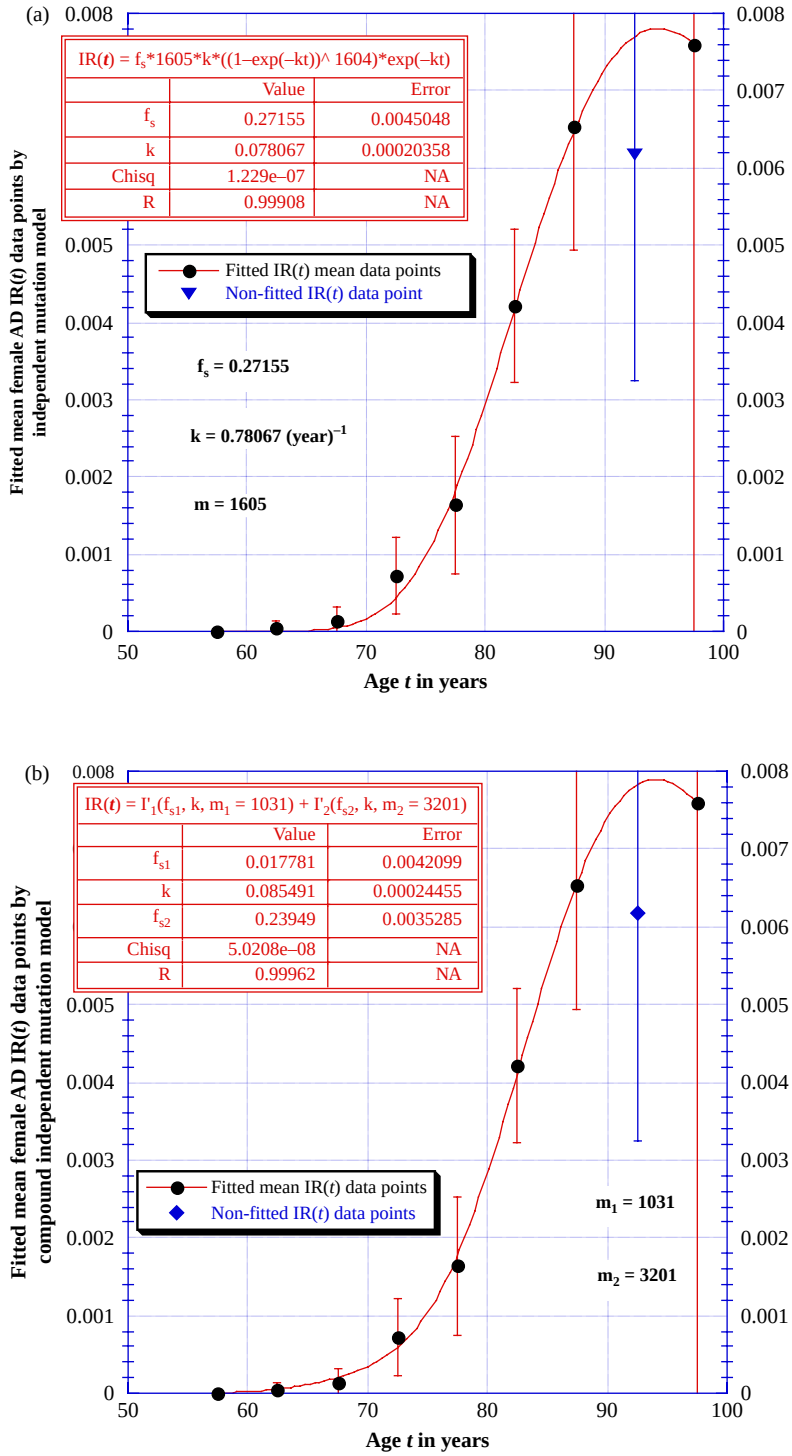


Figure 6. (a) Female AD incidence rate IR(t) data and independent mutation model fit. (b) Compound independent mutation model fit to female AD incidence rate IR(t) data.

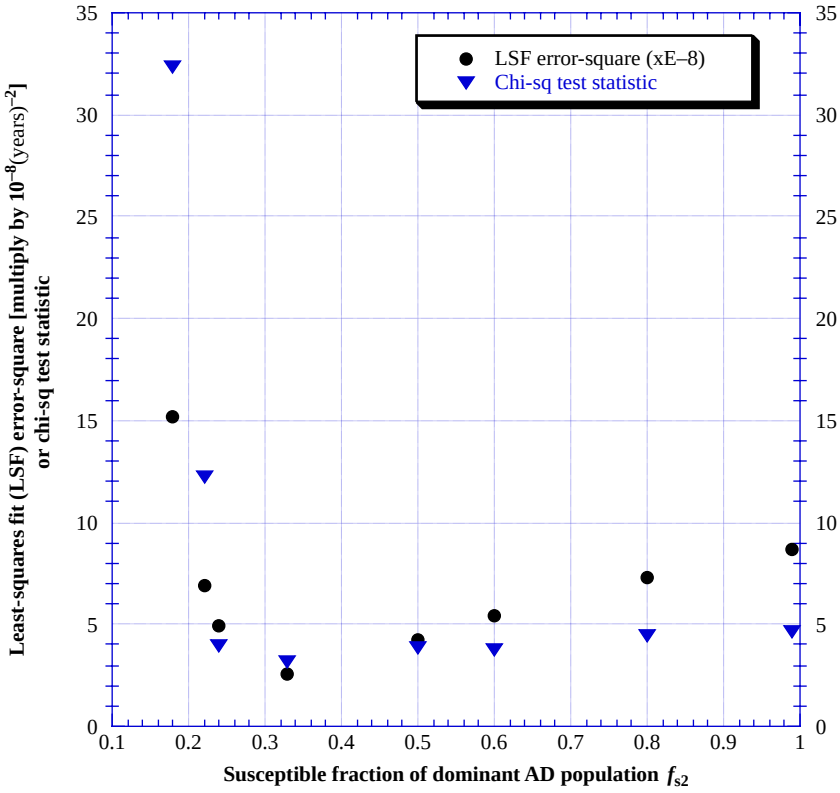


Figure 7. The values of chi-sq test statistic and least-square fit error-square as functions of the value of the susceptible fraction f_{s2} of the dominant AD-susceptible population.

Taking the average age of the over 85-year-old bracket to be 95 years, the average prevalence at age 85 years from the same tables is (1/2)[0.02897 + 0.03376] = 0.03136. Thus, the number of AD cases in this age bracket is expected to rise by 1,297,000 × 0.03136 = 40,680 cases by the year 2007. Thus, the increase in the number of AD patients in the year 2007 over the result in the year 2000 is estimated to be 12,797 and 40,680 for these two age brackets, respectively, for a total increase of 53,477 cases. Thus, the total number of AD patients in the USA in the year 2007 is estimated to be 504,877 (451,400 + 53,477), or 0.176% of the projected population (growing but still quite small).

By the year 2020 the 65–85-year-old population in the USA is projected to increase by 14,000,000 over what it is in 2007; the number of new AD patients this increase is expected to add to the 2007 total is about 76,958. Also by the year 2020, the 85 + year-old population in the USA is projected to increase by 1,705,000 over what it is in 2007; the number of new AD patients this increase is expected to add to the 2007 total is about 53,468. Thus, by the year 2020, the number of AD patients in the USA is projected to reach about 635,297 cases, an increase of 23.8% over what it is in 2007. Thus, the aging of the USA population will result in a significant increase in the number of AD patients.

8. What causes AD and can the disease be blocked?

The average brain contains about 130 billion neurons which are cells that transmit electro-chemical signals between the brain and the nervous system. Although there are many types

of neurons, varying in diameter from 4 to 100 μm and length from less than an inch to several feet, they all have dendrites and axons which allow them to be electrically connected to each other. The synapse is the gap between the axon of one cell and the dendrite of another. A typical neuron can communicate with 1000–10,000 other neurons. The death of a neuron associated with memory destroys all the electrical connections the cell had with the other neurons to which it communicated and is the immediate cause of the onset of AD.

Most of the brain actually consists of glial cells that, unlike neurons, do not carry impulses. Glial cells serve a support function in that they digest parts of dead neurons, manufacture myelin for the neuron coat, and provide physical and nutritional support for neurons. The star-shaped astrocytes, a type of glia cell, are of notable importance since they promote the formation and regulation of synapses without which there would be no memory. For example, astrocytes remove excitatory chemicals from the synapses thereby preventing over-stimulation that can lead to seizures. Astrocytes also form the blood–brain barrier that prevents many toxic substances from crossing into the brain from the blood, another particularly important function since defects in astrocyte function can lead to mutation-causing antigens crossing into the brain. Both neurons and astrocytes are made from the same stem cell, and the ratio of these cells must be carefully regulated for the brain to function properly. A key receptor on the stem cell, erbB4, controls the production of astrocyte cells, but the function of this receptor is controlled, in turn, by presenilin which plays a major role in AD [5]. Genetic mutations in the presenilin gene lead to defective variations of this protein which leads to astrocyte malfunction, a weakening of the blood–brain barrier, synaptic malfunction, and AD.

A loss of brain neurons can be fatal. For every minute delay in the treatment of an average acute stroke the brain losses 1.9 million neurons, 13.8 billion synapses, and 7.4 miles of myelinated fibres [6]. Compared with the rate of neuron loss in normal brain aging, the blood-starved brain ages 3.6 years each hour without treatment [6]. At the above rates, if such a stroke runs its full course by being left untreated for an average of 10 h, the brain loses 1.2 billion neurons, 8.3 trillion synapses, and 4470 miles of myelinated fibres [6]. Thus, stroke data supports another assumption of the model developed here, namely the continual destruction of neurons in AD eventually leads to death.

Pakkenberg and Gundersen stereographically analysed the brains of 94 healthy Danish citizen: 62 males, ages range 19–87 years, and 32 females, age range 18–93 years [7]. The average numbers of neocortical neurons in the study females and males was 19 billion and 23 billion, respectively, a 16% difference. However, both sexes were found to have lost about 10% of their neocortical neurons over the period from the age of 20 years to the age of 90 years. Thus, it is normal for the brain to lose about 85,000 neocortical nerve cells per day. This normal neuron loss is not associated with AD and will be assumed to occur in such an ordered way so as to prevent the symptoms of AD.

A variety of brain cell genes have been linked to AD, but the most important are those associated with the proteins presenilin 1, presenilin 2 and amyloid precursor protein (APP). More than 40 different missense mutations in the presenilin 1 gene and 6 different mutations in the presenilin 2 gene have been linked to AD. APP is necessary for normal brain function, but APP can be abnormally cleaved by enzymes to form amyloid beta peptide, a substance associated with AD.

One of the characteristics of AD is the abnormal build-up of extracellular amyloid plaque in the cerebral cortex of the brain which has been linked to mutations in the genes encoding APP. Six different mutations of the APP gene have been linked to AD. A popular model of AD development is that hypothesis that the accumulation of extracelluar amyloid

plaque triggers a pathogenic cascade leading to neurodegeneration and the disease [8]. However, *amyloid plaque accumulation is not a necessary feature of dementia or neuorodegeneration*. The distribution of synaptic loss correlates well with the pattern of dementia and severity of dementia in AD *but the amyloid plaque distribution does not* [9]. Equally interesting, mouse models duplicating the over-expression of human mutant amyloid plaque production have largely failed to reproduce neurodegeneration [10].

Jie Shen and Raymond Keller have advanced the plausible hypothesis that AD is entirely caused by mutations in the presenilin genes [11]. Inactivating presenilins in the cerebral cortex of knockout mice leads to the age-related neurodegeneration characteristic of AD, including synaptic loss, neuronal cell death, and astrogliosis, suggesting that presenilin mutations alone are entirely responsible for AD. In this model, the increase in extracellular amyloid plaque interferes with presenilin function which, in turn, causes AD.

Thus, the mutation model used to calculate the prevalence of AD with age is completely consistent with the current bio-medical research literature on how the disease develops. Clearly, then, the development of AD can be prevented if the causes of the mutations in brain cells that lead to the disease can be blocked. To date, the causes of the mutations in brain cells that lead to AD are unknown. However, identifying these causes could lead to a way of at least delaying the onset of this disease if not blocking its development outright.

9. Testing the ordered mutation model

A stereological analysis of the neocortical cell population of the brain in a cohort of 14 AD patients and a group of 20 normal controls was undertaken by Pelvig *et al.* [12] in an effort to quantify the cell loss due to AD. For the AD group, the *mean* total number of neocortical glial cells and neurons was found to be 25.9×10^9 and 18.9×10^9, respectively [12]. Thus, the ratio of the mean number of glial cells to neurons in the AD group is 1.37. For the control group, the *mean* total number of neocortical glial cells and neurons was found to be 29.1×10^9 and 21.2×10^9, respectively [12]. Thus, the ratio of the mean number of glial cells to neurons in the control group is also 1.37. From these ratios alone it would appear that the mean brain of the AD cohort and that of the controls were scaled versions of each other, with the mean AD brain about 11% smaller than the mean brain of the controls. Although it is tempting to ascribe this 11% difference in mean brain sizes to the disease, this conclusion would only be possible if the controls were an excellent match for the AD cohort in the Pelvig study. Remember that in the Pakkenberg study of *healthy* brains it was found that the average number of neocortical neurons in females and males was 19×10^9 and 23×10^9, respectively, a 16% difference [7]. This same study also found that both sexes lost about 10% of their neocortical neurons over the period from the age of 20 years to the age of 90 years [7]. Thus, sex and age influence the average brain size. Notice that the average brain of the *healthy* females in the Pakkenberg study was about as large as the mean brain of the AD cohort in the Pelvig study. Thus, a smaller brain is not *by itself* an indication of disease.

Is the control cohort in the Pelvig study [12] a good match to the AD cohort? In general, the size of a brain scales with the person's weight, *i.e.* the size of the brain is directly proportional to the person's weight on the average. Of the 14 AD cases included in the Pelvig study, 7 or half of the total had unknown heights and weights; this was also true of 2 of the 20 controls. Thus, the average weight of the *entire* AD cohort in [12] is simply not known. The average mass of the remaining seven AD cases was 56 kg, while that of the

Table 6. Fractional neocortical neuronal loss in the 14 AD patients in study [12] computed from the compound ordered mutation AD model interpretation of the results in Tables 1(b1), (b2) and 2(b1), (b2).

Neocortical neuron loss in AD: Pelvig study patients [12]					
Males			Females		
Age t in years at death	Number	Fractional neuronal loss from model: $P(t)$	Age t in years at death	Number	Fractional neuronal loss from model: $P(t)$
68	1	0.001461	73	1	0.002681
81	1	0.01369	74	1	0.003450
90	1	0.04862	75	1	0.004422
91	1	0.05478	80	2	0.01379
			81	2	0.01686
			82	1	0.02041
			86	1	0.03979
			94	1	0.09858
Average fractional neocortical neuronal loss for 14 AD patients: 0.01571 (1.571%)					

remaining 18 controls was 61.55 kg (see Table 1 in [12]). Thus, the known weights of the control cohort was 9.9% larger than the known weights of the AD cohort on the average, and this could account for most of the 11% measured difference in average brain size between these two cohorts.

Although lighter people have smaller brains than heavier people all things being equal, *they generally are not*. A cohort of people of the same sex, age and weight will nonetheless have a distribution or spread of brain sizes (unmeasured to date) with an accompanying standard deviation (also unknown) because of *genetic* factors, making finding a good control cohort for a given AD cohort even harder.

Thus AD is one of many factors that influence brain size. Nonetheless, the results of the Pelvig study [12] suggest that there may be *some* brain cell loss due to AD, and this loss supports one of the assumptions in the mutation models discussed in this paper.

The modelling in this paper can in fact be used to quantify the percentage of brain cell loss for the AD cohort in the Pelvig study *that is solely due to the disease*. This estimate must assume that the AD cohort and controls in [12] were the *average* cases considered in the modelling in this paper.

Using the model results in Tables 1(b1),(b2) and 2(b1),(b2) for the fraction of neurons lost with age in males and female AD patients, respectively, on the average, the fractional neocortical neuron lost for the patients in the Pelvig study [12] can be computed, and the result is shown in Table 6. Thus, the model predicts that the average cortical neuronal loss for Pelvig's AD cohort was 1.57%. This percentage is small enough to be overwhelmed by sex, age, and genetic differences between the AD and control cohorts, and indeed, other studies similar to Pelvig's failed to find any significant difference between these cohorts.

10. Remaining survival time following a diagnosis of AD is reduced

Brookmeyer *et al.* [13] found that a diagnosis of AD at any age invariably shortens the remaining survival time of the patient on the average. In the Baltimore Longitudinal Study

Table 7. Calculation of chi-square goodness-of-fit test statistic $\chi^2 = \sum_1^8 \chi_j^2$ for the compound ordered model fit to male AD data shown in Figure 3(c).

Age interval (in years)	60–64 ($j = 1$)	65–69 ($j = 2$)	70–74 ($j = 3$)	74–79 ($j = 4$)	80–84 ($j = 5$)	85–89 ($j = 6$)	90–94 ($j = 7$)	95–99 ($j = 8$)	Row Sum
Observed frequency, O_j	6	12	20	27	34	29	13	1	130
Sample population, n_j	8889	6611	5101	3837	2111	1052	312	66	
Initial prevalence of interval, p_i	2.00×10^{-4}	7.58×10^{-4}	2.16×10^{-3}	5.22×10^{-3}	1.17×10^{-2}	2.49×10^{-2}	4.86×10^{-2}	8.37×10^{-2}	
Final prevalence of interval, p_f	7.58×10^{-4}	2.16×10^{-3}	5.22×10^{-3}	1.17×10^{-2}	2.49×10^{-2}	4.86×10^{-2}	8.37×10^{-2}	1.26×10^{-1}	
Model prediction of θ_j for interval	5.58×10^{-4}	1.40×10^{-3}	3.07×10^{-3}	6.55×10^{-3}	1.34×10^{-2}	2.45×10^{-2}	3.76×10^{-2}	4.76×10^{-2}	
Expected frequency (raw), $\theta_j \cdot n_j$	4.96	9.30	15.6	25.1	28.4	25.8	11.7	3.14	124
Expected frequency normalized, E_j	5.19	9.73	16.3	26.2	29.7	27.0	12.2	3.28	130
$\chi_i^2 = (O_j - E_j)^2/E_j$	0.125	0.525	0.795	0.019	0.600	0.141	0.043	1.59	3.838

Table 8. Calculation of chi-square goodness-of-fit test statistic $\chi^2 = \sum_1^8 \chi_j^2$ for the compound ordered model fit to female AD data shown in Figure 5(c).

Age interval (in years)	60–64 ($j = 1$)	65–69 ($j = 2$)	70–74 ($j = 3$)	74–79 ($j = 4$)	80–84 ($j = 5$)	85–89 ($j = 6$)	90–94 ($j = 7$)	95–99 ($j = 8$)	Row Sum
Observed frequency, O_j	3	7	35	66	124	109	42	12	398
Sample population, n_j	11037	10340	9785	8072	5749	3170	1234	267	
Initial prevalence of interval, p_i	1.27×10^{-4}	3.89×10^{-4}	1.25×10^{-3}	4.42×10^{-3}	1.37×10^{-2}	3.41×10^{-2}	6.68×10^{-2}	1.06×10^{-1}	
Final prevalence of interval, p_f	3.89×10^{-4}	1.25×10^{-3}	4.42×10^{-3}	1.37×10^{-2}	3.41×10^{-2}	6.68×10^{-2}	1.06×10^{-1}	1.44×10^{-1}	
Model prediction of θ_j for interval	2.62×10^{-4}	8.68×10^{-4}	3.17×10^{-3}	9.45×10^{-3}	2.08×10^{-2}	3.44×10^{-2}	4.35×10^{-2}	4.30×10^{-2}	
Expected frequency (raw), $\theta_j \cdot n_j$	2.89	8.97	31.0	76.3	120	109	53.7	11.4	413
Expected frequency normalized, E_j	2.78	8.63	29.8	73.4	115	105	51.7	11.0	398
$\chi_i^2 = (O_j - E_j)^2/E_j$	0.016	0.310	0.880	0.748	0.633	0.146	1.825	0.082	4.640

Table 9. Calculation of chi-square goodness-of-fit test statistic $\chi^2 = \sum_1^8 \chi_j^2$ for the compound independent model fit to male AD data for $f_{s2} = 0.0329$ (see point in Figure 7).

Age interval (in years)	60–64 ($j = 1$)	65–69 ($j = 2$)	70–74 ($j = 3$)	74–79 ($j = 4$)	80–84 ($j = 5$)	85–89 ($j = 6$)	90–94 ($j = 7$)	95–99 ($j = 8$)	Row Sum
Observed frequency, O_j	6	12	20	27	34	29	13	1	130
Sample population, n_j	8889	6611	5101	3837	2111	1052	312	66	
Initial prevalence of interval, p_i	1.94×10^{-4}	9.22×10^{-4}	2.78×10^{-3}	6.50×10^{-3}	1.40×10^{-2}	2.94×10^{-2}	5.61×10^{-2}	9.38×10^{-2}	
Final prevalence of interval, p_f	9.22×10^{-4}	2.78×10^{-3}	6.50×10^{-3}	1.40×10^{-2}	2.94×10^{-2}	5.61×10^{-2}	9.38×10^{-2}	1.38×10^{-1}	
Model prediction of θ_j for interval	7.28×10^{-4}	1.86×10^{-3}	3.74×10^{-3}	7.62×10^{-3}	1.57×10^{-2}	2.79×10^{-2}	4.07×10^{-2}	4.99×10^{-2}	
Expected frequency (raw), $\theta_j \cdot n_j$	6.47	12.3	19.0	29.2	33.1	29.4	12.7	3.29	145.7
Expected frequency normalized, E_j	5.77	11.0	17.0	26.0	29.5	26.2	11.3	2.94	130
$\chi_i^2 = (O_j - E_j)^2/E_j$	0.009	0.09	0.521	0.032	0.657	0.293	0.241	1.28	3.124

of Aging cohort, a diagnosis of AD at the age of 65 years reduced the remaining survival time by 67% in both men and women [13]; a diagnosis of AD at the age of 90 years reduced the remaining survival time by 37% in men and 42% in women. The disease model presented here can readily explain the shortening of the average remaining survival time by a diagnosis of AD.

From (A6) in the appendix the ratio $N_m(t)/N_0$ [with $f_s = 1$] is the fraction of the population that has developed AD at age t. This ratio also represents the fraction of an average person's brain cells that have died at age t. Since $N_m(t)/N_0$ is a monotonically increasing function, the fraction of an average person's brain cells that have been destroyed at age 65 years is considerably less than that at age 90 years. Clearly, late AD can be developed at any age as long as a non-zero fraction of a person's brain cells have been destroyed.

From Table 1(b1) and the modelling in this paper, only about 0.0677% of the cells of an average male diagnosed with AD at the age of 65 years have been destroyed while about 5.55% of the cells of an average male diagnosed with AD at the age of 90 years have been destroyed. Clearly, the 0.0677% of brain cells that have been destroyed in men that develop AD at the age of 65 years are just as important insofar as AD is concerned as the 5.55% of brain cells that have been destroyed in men that develop AD at the age of 90 years. Thus, not all brain cells are equally important insofar as AD is concerned. Since people diagnosed with AD at the age of 65 years have over 99.9% of their brain cells in tact they tend to live longer than people diagnosed with AD at the age of 90 years who have lost over 5% of their brain cells. In the model presented in this paper, death from AD can be viewed as resulting from the destruction of a sufficient number of brain cells controlling the central nervous system; this takes longer to occur at age 65 years than at age 90 years. Thus, the model presented here offers a straight-forward, plausible explanation of the empirical results in Ref. [13].

11. Conclusion

The modelling in this paper suggests that there may be widespread natural immunity to the development of AD. Biomedically determining the mechanism of this immunity should become a research priority since such knowledge could help in discovering methods to block AD from developing altogether.

The number of people in the USA estimated to have AD in the year 2000 was 451,000, only 0.16% of the population at that time. Although this number is expected to grow as the population ages, it nonetheless is far less than other estimates that have appeared in the popular press.

Acknowledgement

Supported in part by UMBC Interdisciplinary Research Grant (2006).

References

[1] E. Kokmen, C.M. Beard, P.C. O'Brien, K.P. Offord, and L.T. Kurland, *Is the incidence of dementing illness changing?*, Neurology 43 (1993), pp. 1887–1892.

[2] W.A. Rocca, R.H. Cha, S.C. Waring, and E. Kokmen, *Incidence of dementia and Alzheimer's disease, a reanalysis of data from Rochester, Minnesota, 1975–1984*, Am. J. Epidemiol. 148(1) (1998), pp. 51–62.

[3] C. Kawas, S. Gray, R. Brookmeyer, J. Fozard, and A. Zonderman, *Age-specific incidence rates of Alzheimer's disease: The Baltimore longitudinal study of aging*, Neurology 54 (2000), pp. 2072–2077.

[4] L.E. Hebert, P.A. Scheer, L.A. Beckett, M.S. Albert, D.M. Pilgrim, M.J. Chown, H. Funkelstein, and D.A. Evans, *Age-specific incidence of Alzheimer's disease in a community population*, JAMA 273(17) (1995), pp. 1354–1359.
[5] S.P. Sardi, J. Murtie, S. Koirala, B.A. Patten, and G. Corfas, *Presenilin-dependent ErbB4 nuclear signaling regulates the timing of astrogenesis in the developing brain*, Cell 127(1) (2006), pp. 185–197.
[6] J.L. Saver, *The presenilin hypothesis of Alzheimer's disease. Evidence for a loss-of-function pathogenic mechanism*, Stroke 37 (2006), pp. 263–266.
[7] B. Pakkenberg and H.J.G. Gundersen, *Neocortical number in humans: Effect of sex and age*, J. Comp. Neurol. 384 (1997), pp. 312–320.
[8] J. Hardy and D.J. Selkoe, *The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics*, Sci. Tech. Froid 297(5580) (2002), pp. 353–356.
[9] R.D. Terry, E. Masliah, D.P. Salmon, N. Butters, R. DeTeresa, R. Hill, L.A. Hansen, and R. Katzman, Ann. Neurol. 30 (1991), pp. 572–580.
[10] M.C. Irizarry, M. McNamara, K. Fedorchak, K. Hsiao, and B.T. Hyman, J. Neuopathol. Exp. Neurol. 56 (1997), pp. 965–973.
[11] J. Shen and R.J. Keller, PNAS 104(2) (2007), pp. 403–409.
[12] D.P. Pelvig, H. Pakkenberg, L. Regeur, S. Oster, and B. Pakkenberg, *Neocortical glial cell numbers in Alzheimer's disease: A sterological study*, Dement. Geriatr. Cogn. Disord. 16 (2003), pp. 212–219.
[13] R. Brookmeyer, M.M. Corrada, F.C. Curriero, and C. Kawas, *Survival following a diagnosis of Alzheimer's disease*, Arch. Neurol. 59 (2002), pp. 1764–1767.
[14] N.A. Weiss and M.J. Hassett, *Physical basis of cognitive alterations in Alzheimer's disease: Synapse loss is the major correlate of cognitive impairment*, in: *Introductory Statistics*, 2nd ed., Addison-Wesley, Reading, MA, 1988.

Appendix

1. The ordered mutation model of AD

In this model the onset of AD will be assumed to require an *ordered* chain of m mutations to occur within a neuron leading to its death, where m is a positive integer. Suppose in the average human brain in a risk population there are a total of N_T neurons, any one of which could trigger the onset of AD if it dies. Suppose further that m ordered mutations are required in a neuron for it to die. At any given age, the number of neurons that have experienced q mutations will be denoted by $N_q(t)$, where $q \equiv 0, 1,2, \ldots, m$. Thus, in this notation, the number of neurons that have experienced no mutations at age t is denoted by $N_0(t)$, and the number of neurons that have died by age t is denoted by $N_m(t)$. Figure 1 is a schematic representation of the ordered mutation model. The m mutation rates (constants) k_r, $r \equiv 1,2, \ldots, m$, shown in Figure 1 are defined as the fraction of cells in mutation state $r-1$ that undergo the rth mutation per unit time. This mutation model is *ordered* so that only cells in the $r-1$ mutation state can mutate into the rth mutation state, and cells in the $q-1$ state cannot mutate into any other mutation state as indicated in Figure 1.

It is possible that a fraction of the total number of neurons N_T are actually *immune* to undergoing the m mutations necessary for them to die and contribute to AD. If this immune fraction is denoted by f_i, then the fraction of neurons susceptible to mutating to death is $f_s = 1 - f_i$, where $0 \leq f_s \leq 1$. Thus, in a total of N_T neurons the number susceptible to contributing to the development of AD is given by $N_s = f_s N_T$.

The set of coupled, time-dependent equations involving the neuron numbers $N_r(t)$ at age t in this model are

$$\frac{dN_0(t)}{dt} = -k_1 N_0(t), \tag{A1a}$$

$$\frac{dN_1(t)}{dt} = k_1 N_0(t) - k_2 N_1(t), \tag{A1b}$$

$$\frac{dN_s(t)}{dt} = k_s N_{s-1}(t) - k_{s+1} N_s(t), \quad s \equiv 1, 2, 3, \ldots, m-1, \tag{A1c}$$

$$\frac{dN_m(t)}{dt} = k_m N_{m-1}(t). \tag{A1d}$$

Adding these equations together and integrating leads to

$$N_S = N_0(t) + N_1(t) + N_2(t) + N_3(t) + \cdots + N_m(t) \equiv \sum_{p=0}^{m} N_p(t), \tag{A2}$$

since $N_r(0) = 0$ for $r \equiv 1, 2, \ldots, m$.

The solutions to each of the Equations in (A1) *in order* are

$$N_0(t) = N_s e^{-k_1 t} = f_s N_0 e^{-k_1 t}, \tag{A3a}$$

$$N_s(t) = e^{-k_{s+1} t} \int_0^t k_s N_{s-1}(t') e^{k_{s+1} t'} dt', \quad s \equiv 1, 2, \ldots, m-1, \tag{A3b}$$

$$N_m(t) = k_m \int_0^t N_{m-1}(t') dt'. \tag{A3c}$$

If $(k_r,t) \ll 1$ for $r = 1,2, \ldots, m$, then the Equations in (A3) yield the approximate solution

$$N_m(t) \approx f_s N_0 k_1 k_2 k_3 \ldots k_m \frac{t^m}{m!}, \quad (kt) \ll 1. \tag{A4}$$

Thus, the value of m, the number of mutations necessary to cause the onset of AD, can be determined by fitting the function in (A4) to the *early part* of the $N_m(t)$ data curve.

If all the mutation rates k_r, $r = 1,2,\ldots,m$, are equal to the same constant k, then the solutions in (A3b) yield the particularly simple solutions

$$N_s(t) = f_s N_T \frac{k^s}{s!} t^s e^{-kt}, \quad s \equiv 1,2,\ldots,m-1, \tag{A5a}$$

and the number of neurons that have developed AD by time t is given by

$$N_m(t) = f_s N_T \left[1 - e^{-kt} \sum_{s=0}^{m-1} \frac{(kt)^s}{s!} \right]. \tag{A5b}$$

Notice that $N_m(0) = 0$, as it should, and $N_m(\infty) = N_s = f_s N_T$ so that the entire collection of susceptible neurons die as t becomes infinite, as they must.

From the result in (A5b) the probability $P(t)$ that a neuron will undergo the ordered set of m mutations and die at age t is given by

$$P(t) = \frac{N_m(t)}{N_T} = f_s \left[1 - e^{-kt} \sum_{s=0}^{m-1} \frac{(kt)^s}{s!} \right]. \tag{A6}$$

The probability in (A6) is also equal to the probability of developing AD at age t in the risk population, a quantity also known as the prevalence of AD at age t. Thus, in a population of a total number of H_T humans, if the number of humans that have been diagnosed with AD by age t is denoted by $H_{AD}(t)$, then (A6) leads to

$$P(t) = \frac{H_{AD}(t)}{H_T} = f_s \left[1 - e^{-kt} \sum_{s=0}^{m-1} \frac{(kt)^s}{s!} \right] \equiv \int_0^t IR(t')dt'. \tag{A7}$$

Using (A7), the fraction of the population that comes down with AD per unit time at age t (the *fractional AD incidence rate*) is given by

$$IR(t) \equiv I(f_s, k, m) \equiv \frac{dP(t)}{dt} = \frac{1}{H_T} \frac{dH_{AD}(t)}{dt} = \frac{f_s k^m}{(m-1)!} t^{(m-1)} e^{-kt}, \tag{A8}$$

where m is the number of mutations necessary to cause AD, k is the mutation rate, and f_s is the fraction of the population that is susceptible to developing AD. The values of the three parameters m, k and f_s are determined by fitting the incident rate function in (A8) to AD incidence data. If every member of the population is susceptible to developing AD, then $f_s = 1$ and the fit involves determining the values of only two parameters, m and k.

An inherent feature of the model developed here is that the cumulative number of people in a risk population that develop AD over time can never exceed the total number of people in the risk population, a characteristic known as *saturation*. The saturation of the model is responsible for the fact that the general AD incidence function in (A8) monotonically increases, peaks, and monotonically declines towards zero as the age of the risk population increases. The peak in the incidence function in (A8) occurs at the age of

$$t_{max} = \frac{(m-1)}{k}. \tag{A9}$$

The following procedure was followed in obtaining the least-squares fits shown in this paper. Referring to (A8), the parameter

$$b \equiv \frac{f_s H_T k^m}{(m-1)!} \tag{A10}$$

was regarded as an independent parameter to be determined, along with m and k, by the fit. Then, solving (A10) for f_s, the value of f_s can be calculated from the values of b, m and k returned by the fit.

If the set of n consecutive data points used in the fit are denoted by $\{d_i\}$ and if the corresponding model fit for this points are denoted by the set $\{x_i\}$, then the square of the error of the fit, to be called *chisq*, is defined as

$$\text{Chisq} \equiv \sum_{i=1}^{n} [x_i - d_i]^2. \tag{A11}$$

The lower the value of chisq, the better the model fit is to the data.

If $kt \ll 1$, then the leading term in the solution in (A5b) is

$$N_m(t) = f_s N_0 \frac{(kt)^m}{m!}, \quad (kt) \ll 1. \tag{A12}$$

Equating (A4) and (A12) leads to the following relationship between the average mutation rate k and the set m mutation rates $k_1, k_2, \ldots, k_m$:

$$k^m = k_1 k_2 k_3 \ldots k_m. \tag{A13}$$

The *physical* model that produces the result in (A8) for the AD incidence rate has distinct advantages over the *phenomenological* models whose independent parameters generally lack a simple, clear physical interpretation.

If all the mutation rates have different values, then the time-dependent AD incidence rate can be calculated in a straightforward way using the results in Equations (A3).

For example, if $m = 1$, then using (3a) and remembering that for $r = 1, 2, \ldots, m$

$$\text{IR}(1; k_r; t) \equiv \frac{dN_1(t)}{dt} = k_1 N_0(t) = f_s N_0 k_1 e^{-k_1 t}, \tag{A14a}$$

and

$$N_1(t) = f_s N_0 [1 - e^{-k_1 t}]. \tag{A14b}$$

Similarly, if $m = 2$, then

$$N_1(t) = f_s N_0 \frac{k_1}{(k_2 - k_1)} [e^{-k_1 t} - e^{-k_2 t}], \tag{A15a}$$

and

$$\text{IR}_2(2; k_r; t) \equiv \frac{dN_2(t)}{dt} = k_2 N_1(t) = f_s N_0 \frac{k_1 k_2}{(k_2 - k_1)} [e^{-k_1 t} - e^{-k_2 t}]. \tag{A15b}$$

As a final example, if $m = 3$, then

$$\text{IR}(3; k_r; t) \equiv \frac{dN_3(t)}{dt} = k_3 N_2(t) = f_s N_0 \frac{k_1 k_2 k_3}{(k_2 - k_1)} \left[\frac{(e^{-k_1 t} - e^{-k_3 t})}{(k_3 - k_1)} - \frac{(e^{-k_2 t} - e^{-k_3 t})}{(k_3 - k_2)} \right]. \tag{A16}$$

Continuing in this way, the AD incidence rate $\text{IR}_m(t)$ for any value of m and arbitrary (positive) values of the mutation rates $k_1, k_2, k_3, \ldots, k_m$, can be computed.

Notice that the AD incidence functions in (A14a), (A15b), and (A16) all approach zero if the age of the cohort becomes large enough, a characteristic feature of *all* AD incidence functions in this model. For $m > 1$, the incidence function must also vanish at age $t = 0$. Thus, for $m > 1$, it is always the case that the incidence function starts out at zero, monotonically grows until it reach a peak, and then monotonically declines towards zero as the age of the risk population continues increasing. However, the peak in the incidence rate function may occur at ages above the natural human life-span; in these cases not everyone in the risk population that is susceptible to developing AD will get it before they die of something else.

For greater clarity, at points in this exposition the function $N_s(t)$ will be denoted by $N(s/m, t)$, $s = 0,1,2, \ldots, m$, where m is the number of mutations necessary to cause AD in a characteristic cohort.

2. The chi-square (χ^2) goodness-of-fit test

The chi-square goodness-of-fit test ascertains whether two distributions can be regarded as identical to each other from a statistical point of view. In this part of the appendix, this test will be used to test the quality of the compound ordered mutation model fits in Figures 3(c) and 5(c). The discussion here will follow the notation in R14 (see chapter 11).

The steps in applying the chi-square goodness-of-fit test to the fit in Figure 3(c) are arranged in Table 7 that follows. The fastest way to illustrate how this test is executed is to explain how each and every line that appears in this table is obtained.

The top row in Table 7 contains the 8 age intervals used in the collection of male AD incidence data in Table 1(a).

The second row in Table 7 is simply the sum of all the numerators in each column in Table 1(a). Thus, this sum is simply the total number of AD cases within each age interval over a 10-year period.

The third row in Table 7 is simply the sum of all the denominators in each column in Table 1(a). Thus, this sum is simply the total sample population within each age interval.

The fourth row in Table 7 is the *model* calculation of the AD prevalence at the beginning of the age interval, and the fifth line is the *model* calculation of the prevalence at the end of the age interval. The values that appear on these lines are taken directly from Table 4. For example, for the 60–64 age interval $P_{\mathrm{i}} = P(60 \text{ years}) = 0.00020017$ and $P_{\mathrm{f}} = P(65 \text{ years}) = 0.00075842$. These values are necessary to compute the next line in Table 7.

The fifth row in Table 7 is the *model predictions* of the 5-year incidence rates that appear in Table 1(a). To compute the values on this line for each interval, Equation (11b) will be inverted so that

$$\theta(t) = -\ln[1 - P(t)]. \tag{14a}$$

Thus, the value of θ_j for each interval is given by

$$\theta_j = \ln\left[\frac{1 - P_i}{1 - P_f}\right]. \tag{14b}$$

For example, for the 60–64 age interval,

$$\theta_1 = \ln\left[\frac{1 - 0.000200}{1 - 0.000758}\right] = 5.58 \times 10^{-4}.$$

The other values of θ_j on this row are computed in the same way.

The sixth row in Table 7 is the *raw* expected frequency for this age interval, *i.e.* the *raw* number of expected AD cases for this interval computed from the model, and is computed by taking the product of θ_j on row six and n_j on row three. The total number of AD cases on this row will have to be normalized to give the same total number of *observed* AD cases on row two (130 cases). This normalization is performed and the result appears on the next row (row eight).

The chi-square goodness-of-fit test statistic for this problem appears on the last row in Table 7. The sum of the elements on this row is the positive definite chi-square test statistic defined by

$$\chi^2 = \sum_{j=1}^{8} \chi_j^2 \equiv \sum_{j=1}^{8} \frac{[O_j - E_j]^2}{E_j}. \tag{15}$$

The conclusion of the analysis in Table 7 is that $\chi^2 = 3.838$ for the fit to the male AD incidence data shown in Figure 3(c).

Since $\sum_{j=1}^{8} O_j = \sum_{j=1}^{8} E_j$, the number of degrees of freedom here must be reduced by one. Thus, since there are eight data point in the fit in Figure 3(c), there are $8 - 1 = 7$ degrees of freedom in this problem. For a problem with 7 degrees of freedom a χ^2-value of $\chi^2 = 3.838$ means that there is a 79.8% probability (the p-value is 0.798) that the expected (model) and observed (data) distributions are statistically identical. Thus, by any conventionally used criterion, the model fit in Figure 3(c) is very good.

In order for this statistical test above to be valid, two general criteria must be met by the expected frequencies (row eight in Table 7).

Firstly, all expected frequencies must be at least 1. From Table 7, this criterion is met.

Secondly, at most 20% of the expected frequencies can be less than 5. It must be noted that although many texts state this rule as '*all* expected frequencies must be at least 5', research by W. G. Cochran shows that this statement of the rule is too restrictive. The results in Table 7 show that the second criterion is also met.

To test the quality of the compound ordered model fit to the female AD incidence data in Figure 5(c), the above chi-square goodness-of-fit analysis will be repeated here. All of the steps required in this calculation appear in Table 8, which has exactly the same form as Table 7 above. Thus, the above explanation of a row in Table 7 is exactly the same for the corresponding row in Table 8.

Since there are again eight data points in the fit in Figure 5(c), there are again $8-1=7$ degrees of freedom in this problem. From Table 8, the value of test statistic for this fit turned out to be $\chi^2 = 4.640$. For a problem with 7 degrees of freedom a χ^2-value of $\chi^2 = 4.640$ means that there is a 70.3% probability (the p-value is 0.7038) that the expected (model) and observed (data) distributions are statistically identical. Thus, once again, by any conventionally used criterion, the model fit in Figure 5(c) is very good.

The χ^2 analysis above for the compound *ordered* mutation model is easily extended to the compound *independent* model. The AD incidence rate function in the compound independent model also consists of two terms described by Equation (13). Here however, each incidence rate term on the right-hand side of Equation (13) integrates into a prevalence function given by Equation (8).

The value of χ^2 as a function of the value of a model parameter determine the degree of reliability of the model's prediction of the parameter's value obtained by the least-squares fit. For example, the value of χ^2 for the fit in Figure 4(b) using the independent model will be described in detail. All of the steps necessary to compute the value of χ^2 in this case are shown in Table 9 and are identical to those shown in Tables 7 and 8 for the ordered model. As seen in Table 9, the value obtained is $\chi^2 = 3.124$, and this point is plotted in Figure 7 as a function of susceptible fraction f_{s2} of the dominant AD population. Fixing the value of this parameter at another value, refitting the male AD data with the compound independent model incidence function, and then recomputing the corresponding value of χ^2 yields the other points shown in Figure 7.

As seen in Figure 7, the minimum values for *both* the χ^2 and least-square fit errors occurs at $f_{s2} = 0.329$, where $\chi^2 = 3.12$ ($p = 0.8733$), and both errors increase in *either* direction as we move away from this value.

The maximum physically permitted value of f_{s2} in the compound model occurs when $f_{s2} = 1 - f_{s1}$ so that the *entire* population is susceptible to acquiring AD. The value for f_{s2} in this case turns out to be $1 - 0.0111 = 0.988$ with $\chi^2 = 4.67$ ($p = 0.699$). Thus, although this value for f_{s2} is statistically less probable than the value of $f_{s2} = 0.329$ above, the difference is not great enough to rule it out. Thus, although the modelling in this paper suggests that immunity to AD *may* exist, *it does not prove it.*

At the other extreme for $f_{s2} = 0.18$ the least-squares fit for the remaining parameters produces $\chi^2 = 32.3$ ($p = 0.00004$), an extremely unlikely result. Thus, values of $f_{s2} < 0.18$ are statistically implausible, and 0.18 can be regarded as a lower bound on the value of f_{s2}.