A Model for Coronary Heart Disease and Stroke with Applications to Critical Illness Insurance Underwriting I: The Model

Angus S. Macdonald,* Howard R. Waters,† and Chessman T. Wekwete‡

Abstract

In Part I we construct a model for the development of coronary heart disease (CHD) or stroke that either incorporates, or includes pathways through, the major risk factors of interest when underwriting for critical illness insurance. Our main purpose is to develop a model that could be used to assess the impact on insurance underwriting of genetic information relevant to CHD and/or stroke. Our model is parameterized using data from the Framingham Heart Study in the United States. In Part II we extend this model to include other critical illnesses, for example, cancers and kidney failure, and describe some applications of the model.

1. Introduction

1.1 Objectives

Genetics and insurance is a matter of concern to insurers, politicians, special interest groups, and society in general. The issue can be summarized briefly in the following question: Should insurers have access to genetic information about applicants for life or health insurance? On the one hand, insurers not having access to such information means that people who know they are at increased risk of disease, and even early death, may be more likely to purchase insurance at a price that does not take account of their genetic profile. On the other hand, if insurers have access to genetic information, some sections of the community may become uninsurable and hence be denied what many people would regard as a basic right. For a full discussion of these issues see Macdonald (2000, 2001).

The underlying motivation for this research, and for earlier work by the authors (see Macdonald, Waters, and Wekwete 2003a, 2003b), has been to develop models that can be used to quantify the financial effects of insurers having, or not having, access to genetic information when making underwriting decisions and hence to inform the broader discussion that already is taking place on this topic. Our earlier work considered breast and ovarian cancers; this research is concerned with coronary heart disease (CHD) and stroke. An important feature of our earlier work was that mutations at two gene locations, BRCA1 and BRCA2, are known to be responsible for a significant minority of cases of these diseases, possibly between 5% and 10%. There are no comparably simple genetic profiles known to contribute significantly to CHD and stroke. In this sense it could be argued that this research is premature. However, we regard it as useful for the following reasons:

a. It is useful to have such models in readiness. The wider debate on access to genetic information is already taking place in the United Kingdom and other countries, discoveries of genetic links to

* Angus S. Macdonald, FFA, PhD, is a Professor in the Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, U.K., e-mail: A.S.Macdonald@ma.hw.ac.uk.

† Howard R. Waters, FIA, FFA, DPhil, is a Professor in the Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, U.K., e-mail: H.R.Waters@ma.hw.ac.uk.

‡ Chessman T. Wekwete, PhD, is a Research Actuary (Life and Health) at Swiss Re—South Africa, P.O. Box 72209, Parkview 2122, South Africa, e-mail: Chessman_Wekwete@swissre.com.

diseases are occurring at an ever increasing rate, and it takes time to develop models to assess the impact of this research.

b. Current knowledge of the genetic links associated with CHD and stroke indicates that these links may be strongest through associated risk factors, in particular hypertension, hypercholesterolemia, and diabetes (see Section 4). Given this, it seems sensible to develop models for CHD and stroke that incorporate pathways through these risk factors.
c. The models we develop can be used to quantify underwriting decisions relating to risk factors for CHD and stroke but not related to any genetic information (see Part II).

In this part we describe and parameterize a model for the development of CHD and stroke that takes account of the major risk factors for these conditions. In Part II we extend this model to include the other events covered by critical illness (CI) insurance, and we describe some applications of the model. A brief discussion of CI insurance is given in Section 1.2.

In Sections 2 and 3 we discuss CHD and stroke. In Section 4 we discuss briefly some of the genetic links to these diseases. Our model is parameterized using data from the Framingham Heart Study. This data set is described in Section 5. In Section 6 we outline the general features of our model and the methodology we have used to estimate the required parameters. Details of the parameterization are given in Sections 8, 9, 10, and 11. The model itself is presented in Section 12.

It has not been possible to include in this paper or in Part II full details of the work underlying these papers. Readers interested in obtaining more information should consult Wekwete (2002).

1.2 Critical Illness Insurance

Critical illness insurance was introduced into the U.K. market in the mid-1980s and has been a rapidly growing sector of business ever since. The product, which is usually sold in addition to a more conventional life product, for example, a whole life, endowment, or term assurance, provides a lump-sum payment on the occurrence within the term of the policy of one of a listed number of events. These events include the diagnosis of one of a specified list of diseases, a specified surgical procedure, or total and permanent disability. The major events covered by CI policies are heart attack, stroke, cancer, kidney failure, coronary artery bypass graft, multiple sclerosis, and total and permanent disability. Between them, these seven events account for around 97% of CI claims in the United Kingdom (Dinani et al. 2000). CI policies can be "stand-alone" or "accelerated benefit." The latter, which are more common in the United Kingdom, pay the sum assured on the earlier of death or one of the listed events; the former pay the sum assured only on the occurrence of one of the listed events. For stand-alone CI policies there is usually a specified period following the event, often 28 days, that the insured life has to survive for the benefit to be paid. In Part II we will consider the extension and application of our model for CHD and stroke to stand-alone CI insurance. CI insurance can be considered the simplest form of insurance to which we can apply our model for CHD and stroke. This is because the benefit is paid on the occurrence of these events, and we do not have to consider the mortality of the insured life after, say, a heart attack as we would in the case of term assurance.

A CHD event, typically a heart attack or the development of CHD to the point where heart bypass surgery is required, or a stroke is of considerable relevance to insurers. Either event may itself be the end point for an insurance policy, for example, CI insurance; it may give rise to a payment from a policy, for example, private medical expenses insurance; or it may be an indicator of possible early payment of benefit from a policy, for example, term or endowment assurance.

In Part II we will be concerned with applicants for (CI) insurance who have not previously suffered a CHD event or a stroke. Such applicants will be assessed at the underwriting stage for the likelihood of either of these events in the future. Major risk factors typically taken into account by underwriters are:

Sex
Age
Body mass index (BMI)

Smoking status
Diabetes status
Cholesterol level and
Blood pressure.

Definitions and information about some of these risk factors are included in Section 7. Some other relatively minor risk factors for CHD and/or stroke, for example, atrial fibrillation and left ventricular hypertrophy, have not been included in our model because we wish to concentrate on major risk factors where genetic links are likely to be, or already have been, found.

An important risk factor not included in the above list is "family history of CHD or stroke." The importance of family history to underwriters indicates that there are genetic and/or environmental factors contributing to CHD and stroke. We discuss this in Section 4.

2. Coronary Heart Disease

Coronary heart disease is a term for a group of disease end points resulting from disorders of the coronary arteries that supply blood to the heart muscle. A significant source of disorders of the coronary arteries is the accumulation of fatty streaks (mainly cholesterol and fat deposits) and the formation of fibrous plaque in the artery walls. This is called atherosclerosis. Atherosclerosis mainly progresses in a gradual manner, usually resulting in a long phase of coronary artery disease without any symptoms. Continued deposits on the artery walls will narrow the arteries themselves and may lead to restrictions in blood flow. However, the plaque may become unstable, leading to the rupture of the plaque lesions. The ruptured lesions interact with the flowing blood, and a clot may be formed. These clots can block the artery or may be carried further by the blood. If they encounter another narrowed section of the arteries, they may cause a severe restriction or even blockage of the blood flow.

The interference in the blood supply to the heart muscle can result in a number of disease end points. The differences in the end points are due mainly to the extent of blood deprivation to the heart muscle and the resultant damage. Angina pectoris occurs when the heart requires more blood than can be supplied by the coronary arteries. It is not associated with muscle damage, but the patient experiences heart pains that can be relieved by resting. Myocardial infarction (MI) occurs when part of the heart muscle dies due to a deficiency in the blood supply. The extreme case is that of sudden death, when the heart fails due to extensive death of the muscle. This can occur within an hour of the onset of the symptoms.

CI policies mainly cover MIs. The following is the definition of heart attack used by one reinsurer in respect of CI policies: "The death of a portion of heart muscle as a result of inadequate blood supply as evidenced by a history of typical chest pain, new electrocardiographic changes and by elevation of cardiac enzymes." We note the importance of the specificity of the definition. The changes described above are permanent, and claimants for CI payments can be tested for the presence of given cardiac enzymes.

3. Stroke

Bamford et al. (1988) define stroke as rapidly developing clinical symptoms and/or signs of loss of focal or global cerebral function, lasting more than 24 hours or leading to death, and without an apparent cause apart from being of vascular origin. This is due to an interruption in the supply of blood to the brain, which is secondary to a primary disease of the heart or blood vessels. Strokes are classified in terms of two types of cerebral damage:

1. Cerebral infarction occurs when there is death of part, or all, of the brain tissue largely due to blood clots or stenosis in arteries blocking the supply of blood to the brain. This is usually called ischaemic stroke and constitutes about 80% of strokes (Gubitz and Sandercock 2000).
2. Cerebral hemorrhage (intracerebral and subarachnoid hemorrhage) is associated with the rupture or break in a blood vessel in the brain. The severity of the resulting stroke depends on the site of rupture

and the volume of blood loss. These are called hemorrhagic strokes and constitute about 20% of all strokes.

Bamford et al. (1988) also define a transient ischaemic attack as an acute loss of focal cerebral or ocular function with symptoms lasting less than 24 hours and due to blood clots or stenosis in arteries blocking the supply of blood.

CI policies define stroke in terms of the occurrence of permanent damage to the brain. The definition of stroke used by one reinsurer in respect of CI policies is: "A cerebrovascular incident resulting in permanent neurological damage. Transient ischaemic attacks are specifically excluded."

4. GENETICS

It is accepted that cardiovascular diseases aggregate in families but do not exhibit the pattern of inheritance shown by single-gene disorders (see Sing and Moll 1990). Hence, it is considered likely that CHD and stroke are multifactorial diseases, so that polymorphisms in a combination of genes, together with environmental factors, contribute to them, possibly through other risk factors, for example, diabetes (see, e.g., Brackenridge and Elder 1998, Chapter 21, Part A).

Various gene locations have been linked to increased likelihood of hypertension, hypercholesterolemia, or diabetes. In general, these studies have not yet progressed to the stage where either prevalence rates for the relevant mutations or penetrance rates for the resulting conditions have been quantified. See Wekwete (2002) for more details and references.

Kardia et al. (1999) report some studies as showing a strong association between the frequency of CHD and the frequency of ApoE genotypes carrying the $\epsilon 4$ allele, but also report that numerous studies have found no such associations.

Familial hypercholesterolemia is an autosomal dominant condition that strongly predisposes to coronary artery disease, which can lead to CHD. This can be caused by a mutation in one of many LDL receptor genes. King, Rotter, and Motulsky (1992) report that the frequency of heterozygotes in most populations that have been studied is around 1 in 500 and that "About 50% of affected heterozygotes develop clinically manifest coronary heart disease by age 50 years in males and 60 years in females." The condition can be detected by measuring cholesterol levels, even at early ages. However, direct DNA detection is not yet feasible in populations other than a few specialized "ancestor" societies because of the multiplicity of mutations causing the condition (see King et al. 1992).

It seems clear that there are currently no simple genetic tests that indicate increased likelihood of CHD or stroke. Kardia et al. (1999) say that "for common complex diseases like Coronary Artery Disease, it is far from clear whether genetic variation will have utility for predicting or treating disease in the population at large." It also seems clear that any such tests discovered by future research may well relate to the known risk factors, hypertension, hypercholesterolemia, and diabetes, rather than to other, as yet unknown, risk factors or directly to CHD and stroke. If and when such research bears fruit, models for the development of CHD and stroke, incorporating pathways through these risk factors, will be useful. This is a major motivational factor for our work.

5. DATA: THE FRAMINGHAM HEART STUDY

To parameterize the model we describe below, we require very detailed data. The best data set for our purposes is that produced by the Framingham Heart Study, a longitudinal study of the adult population of Framingham, a small town near Boston. Details of the study can be found on the National Heart, Lung, and Blood Institute Web site at www.nhlbi.nhi.gov.

Starting in 1949, the original cohort of 2,336 men and 2,873 women, aged between 28 and 62 at that time, from Framingham were given health checks every two years. Anonymized data were available to us from examinations 1 to 20, inclusive, that is, for the period 1949–87 for each individual. The study is ongoing and now includes offspring of the original cohort, but these later data were not available to us at the time of this research.

The data set made available to us provided the following information:

Sex: This was recorded at Examination 1 only.
Smoking: This was recorded as "Yes" or "No" at most, but not all, examinations. "Yes" indicates that the person had been a smoker at some time in the year prior to the examination. For smokers, the average number of cigarettes smoked per day was recorded.

Blood sugar level: This was recorded at all but three examinations.
Blood pressure: At each examination, both systolic and diastolic blood pressure were recorded twice.
Total cholesterol: This was recorded at Examinations 2–10 and 13–15, inclusive.
Weight: This was recorded at all examinations.
Height: This was recorded at Examinations 1, 5, and 13–20, inclusive.

Our parameter estimation methods require values of each of the quantities listed above for each individual still alive at each examination. If one of these values was not recorded at a particular examination, we assumed the last recorded value for that person was still valid.

The data also record for each individual dates of significant events, such as a stroke, MI, or death. CHD events in the data were classified as MI (both recognized and unrecognized), angina pectoris, and coronary insufficiency. To be as consistent as possible with CI definitions, we have taken "CHD event" to be equivalent to MI in the Framingham data. The Framingham data also distinguish between different types of stroke, so we were able to exclude transient ischaemic attacks from our definition (see Section 3).

The quality and extent of this data set is unmatched by any other data of which we are aware. However, although the data were adequate for our purposes, they were not ideal for the following reasons:

a. The data relate to the population of a small town in the United States. We are parameterizing a model that we will assume is relevant to a self-selected population, that is, CI insurance policyholders, in a different geographical region, that is, the United Kingdom.
b. The data relate to the time period 1949–87. We require the parameterization to be valid for the first part of the twenty-first century. This is not a problem if there are no calendar time trends in the data and if we have no reason to believe parameter values derived from the data should not be valid over 40 years later. However, if there are time trends in the data, it may not be clear how to extrapolate such values beyond 1987.
c. Since the data relate to a single closed cohort, we have necessarily fewer data for the younger ages, say, ages 28 to 50, than for the older ages.
d. Since the Framingham study started, high-density lipoprotein (HDL-C), as a proportion of total cholesterol, has been shown to be an important risk factor for CHD. However, only total cholesterol was available to us from the Framingham study, and so we are unable to incorporate HDL as a risk factor in our model.
e. Participants in the Framingham study were healthier than the general Framingham population.

Point (a) is considered in detail in Section 14. Points (b) and (c) are considered in later sections where one or both of them affect the estimation of particular parameters. Point (d) is mentioned in Section 7.4. The bias introduced by point (e) will affect prevalence rates in our data; the observed prevalence rates are likely to be lower than for the general Framingham population. However, our parameterization is based on estimates of *incidence* rates, rather than *prevalence* rates, and these should not be affected by the biased sampling.

6. MODELING

Our model for the development of CHD or stroke for an individual is a continuous time Markov model with a finite state space. Such models have been shown to be extremely useful in insurance contexts. See, for example, Hoem (1988) and Macdonald, Waters, and Wekwete (2003a, 2003b). "Time" in this model is equivalent to the person's age.

The following states in our model are absorbing: "CHD," "Stroke," and "Dead." For our model, "Dead" represents death before a CHD event or a stroke. Since our model will be used in Part II as the basis for CI insurance, both "CHD" and "Stroke" are end points for the model.

The remaining states are all transient and represent different combinations of categories of hypertension, hypercholesterolemia, and diabetes. An important consequence of our choice of model is that we are required to represent each of these risk factors by a finite number of discrete categories, whereas each one in reality is measured on a continuous scale. These categories were determined on the basis of current medical and underwriting practices; we give details in Section 7.

The parameterization of our model involves estimating the intensities of the transitions between the transient states and from the transient states to each of the three absorbing states. These intensities will depend on age, sex, smoking status, body mass index, and the starting state, that is, the particular combination of the categories for the three risk factors. In outline, our methodology is to use occurrence/exposure rates to give point estimates of these intensities, with an assumed Poisson distribution for the number of occurrences, and then to use a generalized linear model (GLM) with a log link to smooth these estimates. For example, let (x, s, sm, w, rf) represent a person who is aged x (exact), is sex s, has smoking status sm, has BMI w, and has a combination of categories of the three risk factors denoted by rf; and let $\lambda^{STR}_{x,s,sm,w,rf}$ denote the intensity of having a stroke for this person. We use the Framingham data to calculate two quantities:

$A^{STR}_{x,s,sm,w,rf}$: the number of strokes (before CHD) among people classified as (x, s, sm, w, rf), and,
$E^{STR}_{x,s,sm,w,rf}$: the total time exposed to the risk of stroke (before CHD) during the period of investigation by people classified as (x, s, sm, w, rf).

For the purpose of these calculations, "age x" means aged between $x - \frac{1}{2}$ and $x + \frac{1}{2}$ exact, where x is an integer. We then assume

$$A^{STR}_{x,s,sm,w,rf} \sim \text{Po}[E^{STR}_{x,s,sm,w,rf} \cdot \lambda^{STR}_{x,s,sm,w,rf}]$$

and

$$\lambda^{STR}_{x,s,sm,w,rf} = \exp(f(x, s, sm, w, rf)),$$

where $f(x, s, sm, w, rf)$ is a linear predictor. (In exceptional cases we assume $f(\,)$ is a product of a function of age and a linear predictor; see Section 8.2.) The linear predictor is determined by a stepwise selection procedure, allowing for possible interactions between the terms and including only those terms found to be statistically significant. In general, a factor is included in a GLM if its inclusion significantly improves the fit of the model, as measured by the reduction in deviance. This reduction in deviance is tested at a 5% probability level. However, some subjectivity has been exercised. No significant factors were ever excluded, but some nonsignificant factors were included in our models. We comment on cases where this occurred below. See Wekwete (2002) for more details of the fitting procedures.

As explained earlier, the three continuously varying risk factors—blood sugar level, total cholesterol, and blood pressure—were used to determine discrete categories for the conditions diabetes, hypercholesterolemia, and hypertension, respectively. At any given examination, each life could be categorized according to either the *current* category or the *highest ever* category reached of each of these conditions. The implication of using the latter rather than the former in relation to blood sugar level, for example, would be that the intensity of a CHD event or a stroke depends more on how high this level has ever been than on its current level. For each of the three risk factors, we tested which of these levels was more informative for the intensity of a CHD event or a stroke by including each one separately and by including each one in the presence of the other in the relevant GLM. The results, not surprisingly, were mixed. In some cases the highest ever level was far more informative than the current level (notably in relation to CHD incidence); in other cases the two were about equally informative (notably in relation to the incidence of stroke). On the basis of these findings, we decided to categorize people according to the highest ever category reached of each of these three conditions. This had the result of

making our Markov model presentationally simpler since "backward" transitions from a higher category of, say, hypertension, to a lower category were not possible. However, the calculation of probabilities and financial functions from the model was not significantly simpler as a result of this decision.

7. COMMENTS ON THE RISK FACTORS

In this section we comment briefly on the risk factors included in our model. In particular, we explain how we have determined the categories for some of these risk factors.

7.1 Body Mass Index

BMI is defined as (weight in kg)/(height in m)^2. We use three categories for BMI, as shown in Table 1. These categories are broadly in line with those suggested by Brackenridge and Elder (1998, Section 18) and Hill and Roberts (1996). They are the categories used in the report on the Health Survey of England 1998 (see Erens and Primatesta 1999). They apply to both sexes and all ages.

7.2 Smoking

Smoking is a risk factor for both CHD and stroke. See Nyboe et al. (1991) and Wolf et al. (1988), respectively. Both these studies indicate that the risk of CHD and stroke for exsmokers is similar to that for nonsmokers irrespective of the time since the former stopped smoking.

7.3 Diabetes

Diabetes is defined in terms of elevated glucose levels in the blood (glycemia). The causes and development of diabetes result in categories of Type 1 diabetes and Type 2 diabetes, among others. Type 1 diabetes is a result of insulin deficiency and typically presents before age 30, although it can occur at any age. The deficiency is due to autoimmune destruction of cells involved in insulin production. Most sufferers will depend on insulin treatment for the rest of their lives. The development of the symptoms is acute, and diagnosis of Type 1 diabetes is often soon after development of the symptoms.

Type 2 diabetes is a result of both insulin resistance and diminished insulin secretion. It typically presents in the 50–65 age group, although it could also present at any age. In contrast to Type 1 diabetes, the hyperglycemia related to Type 2 diabetes develops gradually, resulting in the onset and presence of symptoms going unnoticed for many years.

In accordance with the American Diabetes Association: Clinical Practice Recommendations (2000), we regard someone as being diabetic if their blood sugar level is, or, given the comments at the end of the previous section, ever has been, 126 mg/dL or higher.

7.4 Hypercholesterolemia

Total cholesterol is the sum of low-density lipoprotein (LDL-C), high-density lipoprotein (HDL-C), and triglycerides. High levels of LDL-C and low levels of HDL-C are associated with higher risk of CHD (see Brackenridge and Elder 1998). The relative importance as risk factors for CHD of these three constituent parts of total cholesterol is the subject of ongoing research. We have used total cholesterol as a risk

Table 1
BMI Categories

Range	Category
BMI $\leq$ 25	Normal
25 $<$ BMI $\leq$ 30	Overweight
30 $<$ BMI	Obese

factor in our model since the Framingham data included information on this factor but not on its constituent parts.

Let *chol* denote the (highest ever) recorded value of total serum cholesterol, measured in mg/dL. Based on the National Cholesterol Education Program (2001) categories for total cholesterol, we define three categories of hypercholesterolemia:

Category 0:	$chol < 200$
Category 1:	$200 \le chol < 240$
Category 2:	$240 \le chol$.

The National Cholesterol Education Program is a U.S.-based organization. Some other countries use different guidelines.

7.5 Hypertension

The U.S.-based Joint National Committee on Prevention, Detection, and Treatment of High Blood Pressure (1997) suggests six categories of hypertension. These are determined by the values of systolic and diastolic blood pressure, *sbp* and *dbp*, respectively, and are shown in Table 2. To reduce these to a more manageable four levels, we have combined their "Optimal" and "Normal" categories and their "Hypertension Stage II" and "Hypertension Stage III" categories. We define hypertension levels 0, 1, 2, and 3 to correspond to the Joint National Committee's categories "Optimal or Normal," "High Normal," "Hypertension Stage I," and "Hypertension Stage II and Stage III," respectively. These levels are defined in terms of *sbp* and *dbp* values as shown in Table 3.

For the purposes of our model, individuals are assigned to one of four categories of hypertension following one of the biennial Framingham examinations. This assignment is made as follows. For an individual undergoing Framingham examination number j, let Lev_{1_j} and Lev_{2_j} denote the hypertension level, as determined by Table 3, assigned following the first and second set of blood pressure readings, respectively. (Recall from Section 5 that two sets of blood pressure readings were taken at each of the Framingham examinations.) The hypertension category for this individual following examination number j is Category i, $i = 0, 1, 2$ or 3, where

$$i = \max_{k \le j}(\min(Lev_{1_k}, Lev_{2_k}))$$

so that the "better" of the two sets of readings is always used following an examination, but, as explained in Section 6, we use the highest hypertension category ever reached as a predictor for CHD and stroke.

8. The Intensities of Diabetes, Hypercholesterolemia, and Hypertension

In this section we give details of our models for the intensities of the incidence of the (categories of) the three risk factors, diabetes, hypercholesterolemia, and hypertension. These intensities apply to people

Table 2

Hypertension Diagnosis Guidelines Suggested by the Joint National Committee on Prevention, Detection, and Treatment of High Blood Pressure (1997)

Category	Systolic (mmHg)		Diastolic (mmHg)
Optimal	<120	and	<80
Normal	<130	and	<85
High normal	130–39	or	85–90
Hypertension			
Stage 1	140–59	or	90–99
Stage 2	160–79	or	100–109
Stage 3	≥180	or	≥110

Table 3
Determination of Hypertension Levels

	$sbp < 130$	$130 \leq sbp < 140$	$140 \leq sbp < 160$	$160 \leq sbp$
$dbp < 85$	Level 0	Level 1	Level 2	Level 3
$85 \leq dbp < 90$	Level 1	Level 1	Level 2	Level 3
$90 \leq dbp < 100$	Level 2	Level 2	Level 2	Level 3
$100 \leq dbp$	Level 3	Level 3	Level 3	Level 3

who have not yet had a CHD event or a stroke. They are functions of some, or all, of the following: age, sex, smoking status, BMI, and the (categories of) the other two risk factors.

We assume someone moves between categories of one of the risk factors if their recorded categories are different at two consecutive examinations, in which case we assume the move takes place midway between these examinations. This means that contributions to the exposure and number of cases, as outlined in Section 6, can occur only if we have the relevant information, for example, blood sugar level, recorded at consecutive examinations. This feature does not affect the estimation of the intensities of CHD and stroke since precise dates are given for these events.

8.1 Diabetes

Data on blood sugar level were available to us for each of the following pairs of consecutive examinations: (2, 3), (8, 9), (12, 13), (13, 14), . . . , (17, 18). The incidence rates for diabetes obtained from the Framingham data showed almost no calendar time trends, although some evidence of higher rates at the oldest ages existed in the most recent examinations. We decided to discard the data from examinations 2 and 3 as they were too far displaced in time from the remaining data. Consequently we based our estimation on data from examinations 8, 9, and 12–18.

The GLM retained only age and BMI as significant factors, with the effects of the BMI categories "Normal" and "Overweight" not found to be statistically significantly different. It is interesting to note that the model is the same for the two sexes and that the categories of hypercholesterolemia and hypertension appear to have no significant effect on the incidence of diabetes.

The model for the intensity of diabetes is

$$\lambda_{x,w}^{diab} = \exp(\alpha_{int} + \beta x + \nu_w).$$

Table 4 shows the coefficients for this model.

8.2 Hypercholesterolemia

8.2.1 Incidence of Category 1 Hypercholesterolemia

Figures 1 and 2 show the crude rates of incidence of Category 1 hypercholesterolemia subdivided by age only. The estimates based on data from examinations 2, 3, 4, 5, and 6 and examinations 7, 8, 9, 13, and 14 are shown separately.

Table 4
Coefficients of the Linear Predictor for Incidence of Diabetes

Variable	Coefficient	Value	Std. Error
Body mass index	Intercept(α_{int})	-6.703	3.294×10^{-1}
	Age(β)	4.448×10^{-2}	4.874×10^{-3}
	Normal or overweight(ν_w)	-2.434×10^{-1}	4.332×10^{-2}
	Obese	$-\nu_w$	

Figure 1
Observed Crude Incidence Rates of Category 1 Hypercholesterolemia in Different Time Periods for Males

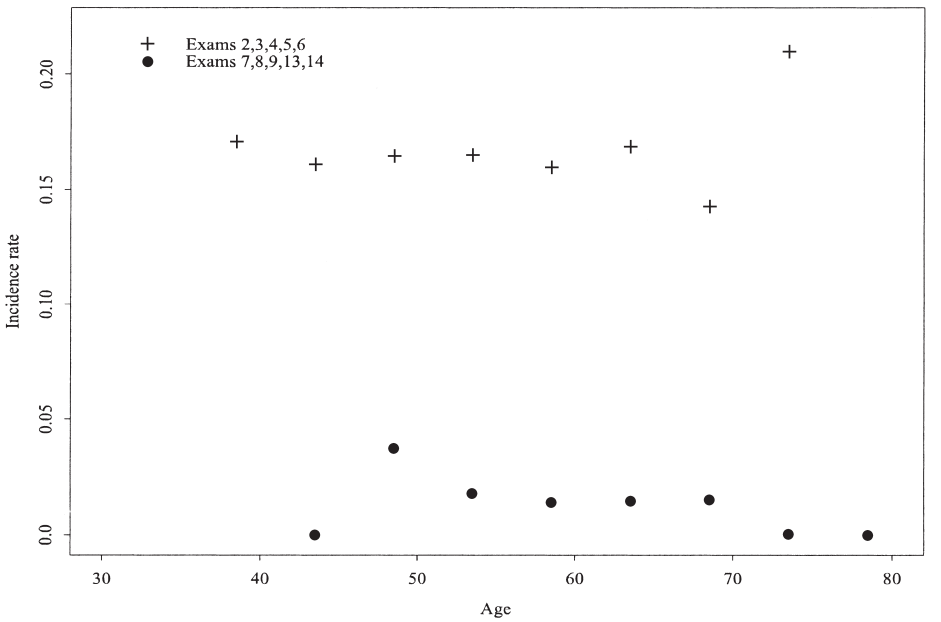
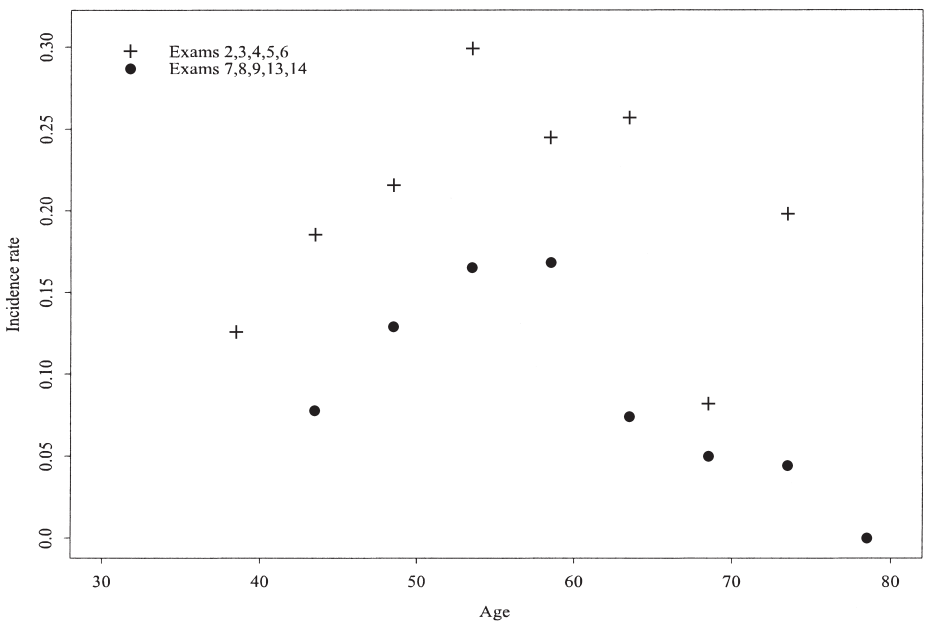


Figure 2
Observed Crude Incidence Rates of Category 1 Hypercholesterolemia in Different Time Periods for Females



Two points stand out from Figures 1 and 2:

1. There is a difference in the levels of incidence for both males and females between the earlier and the later examinations, dramatically so in the case of males. Recall that these two sets of examinations are on average about 12 years apart.
2. For each sex, the shape of the incidence rates as a function of age is roughly the same for the two sets of examinations.

Since it did not seem appropriate to combine the data for the two sets of examinations and since the shapes of the incidence rates were so different for the two sexes, we decided to use the data from the later examinations to parameterize our model for the incidence of Category 1 hypercholesterolemia separately for the two sexes.

The GLM for males retained no factors as significant, not even age, and the resulting model is

$$\lambda_{males}^{chol01} = \exp(-3.312) = 0.036.$$

The modeling of the incidence rates for females was complicated by the fact that the rates clearly depend on age, as can be seen from Figure 2, and the amount of data for younger ages at the later examinations was very limited, as noted in Section 5. To incorporate this age dependence we fitted a function (of age only) to the data for the earlier examinations, with weights inversely proportional to the variance of the estimates. This function is

$$f(x) = \exp(-8.848 + 2.717 \times 10^{-1}x - 2.446 \times 10^{-3}x^2).$$

We then used the data from the later examinations to fit a GLM to the incidence rates with $f(x)$ as an offset function, so that the GLM models the difference between the data and the offset function. The GLM retained no factors as significant and had the form

$$\exp(-0.6446) = 0.5249,$$

which implies that the incidence rates of Category 1 hypercholesterolemia for females are functions of age alone and that the rates for the later examinations can be modeled as a constant multiple (0.5249) of the rates from the earlier examinations. The final model for these incidence rates is

$$\lambda_{females}^{chol01} = \exp(-9.493 + 2.717 \times 10^{-1}x - 2.446 \times 10^{-3}x^2).$$

8.2.2 Incidence of Category 2 Hypercholesterolemia

The incidence rates for Category 2 hypercholesterolemia for both males and females exhibited the same features as the rates for Category 1 hypercholesterolemia for females, as described above. These features were: higher rates for the earlier examinations (multipliers of about 3.8 for males and 2.6 for females), a marked age dependence in the rates and more data, particularly at younger ages, for the earlier examinations. We modeled these incidence rates, separately for males and females, using the same procedure as we had used to model the incidence of Category 1 hypercholesterolemia for females: using weighted least squares to fit a function of age to the rates from the earlier examinations and then incorporating this as an offset when fitting a GLM to the rates from the later examinations. In both cases the resulting GLM retained no factors as significant. The final forms of the models for the incidence rates for Category 2 hypercholesterolemia are

$$\lambda_{males}^{chol12} = \exp(-6.857 + 1.432 \times 10^{-1}x - 1.539 \times 10^{-3}x^2),$$

$$\lambda_{females}^{chol12} = \exp(-15.27 + 4.744 \times 10^{-1}x - 4.470 \times 10^{-3}x^2).$$

The variation over time in the incidence rates for Categories 1 and 2 hypercholesterolemia will be discussed further in Section 14.

8.3 Hypertension

8.3.1 Incidence of Category 1 Hypertension

There was little evidence of a time trend in the data, and so all the available data were used to estimate the incidence of Category 1 hypertension. The GLM retained only age and BMI as significant factors, with the effects of the BMI categories "Normal" and "Overweight" not found to be statistically significantly different. The model is given by the following formula, with coefficients as shown in Table 5:

$$\lambda_{x,w}^{bp01} = \exp(\alpha_{int} + \beta x + \nu_w).$$

8.3.2 Incidence of Category 2 Hypertension

The incidence of Category 2 hypertension was modeled using all the available data. Only age and sex were retained as significant factors. The model is given by the following formula, with coefficients as in Table 6:

$$\lambda_{x,w}^{bp12} = \exp(\alpha_{int} + \beta x + \gamma_s).$$

8.3.3 Incidence of Category 3 Hypertension

The incidence of Category 3 hypertension was modeled using all the available data. As with Category 2 hypertension, only age and sex were retained as significant factors. The model has exactly the same form as that for Category 2 hypertension, as shown above, with coefficients as in Table 7.

9. INCIDENCE OF CORONARY HEART DISEASE

In this section we describe our model for the incidence rates of a CHD event for those people who have not previously had a CHD event or a stroke. The incidence rates are functions of some, or all, of age, sex, smoking status, diabetes, and the categories of BMI, hypercholesterolemia, and hypertension.

The data for males and females exhibited different features, and so we modeled the incidence rates separately for each sex. For both sexes there was no statistical difference between the effects of Categories 0 and 1 for hypertension and between Categories 0 and 1 for hypercholesterolemia so that these two pairs of categories were combined in the models.

For females, both Age and Age^2 were statistically significant, whereas for males, Age^2 was not significant. However, the initial model fitted for males failed to give an incidence rate close to 0 at the youngest ages, around 30. At these ages we would expect the incidence rate for CHD to be very close to 0, but there are very few data here for the reasons explained in Section 5. By introducing an artificially large exposure at age 30, we were able to force our model for males to give incidence rates for CHD close to 0 at these ages. This model did include Age^2 as a statistically significant factor.

With the exception of Age^2 for males, the model retains only those factors that are statistically significant and has the same form for males and for females. In particular, the category of BMI was not statistically significant for either males or females.

The model for the incidence of a CHD event is

Table 5
Coefficients of the Linear Predictor for Incidence of Category 1 Hypertension

Variable	Coefficient	Value	Std. Error
Body mass index	Intercept(α_{int})	-3.969	3.158×10^{-1}
	Age(β)	2.199×10^{-2}	5.231×10^{-3}
	Normal(ν_w)	-9.433×10^{-2}	4.443×10^{-2}
	Overweight or obese	$-\nu_w$	

Table 6
Coefficients of the Linear Predictor for Incidence of Category 2 Hypertension

Variable	Coefficient	Value	Std. Error
	Intercept(α_{int})	-3.865	2.718×10^{-1}
	Age(β)	2.139×10^{-2}	4.331×10^{-3}
Sex			
	Males(γ_s)	-1.300×10^{-1}	3.850×10^{-2}
	Females	$-\gamma_s$	

$$\lambda^{CHD} = \exp(\alpha_{int} + \beta x + \gamma x^2 + \rho_k + \delta_{b^*} + \phi_d + \eta_{c^*}),$$

where the coefficients are given in Table 8 for males and Table 9 for females.

10. INCIDENCE OF STROKE

The data for the incidence rates for stroke indicated that the category of hypercholesterolemia was not a significant factor; that Categories 0, 1, and 2 for hypertension could be combined, so that only the highest category of hypertension had a significantly different effect on the incidence of stroke; and that data for the two sexes could be used in a single model, with sex as a significant factor and with an Age*Sex interaction term. The resulting model for the incidence of stroke is

$$\lambda^{Stroke} = \exp(\alpha_{int} + \beta x + \gamma_s + \rho_k + \delta_{b^*} + \phi_d + \psi x)$$

with coefficients as shown in Table 10.

That cholesterol level is not a significant risk factor for stroke is consistent with other research. In a study of 7,052 men and 8,354 women who had baseline examination in the mid-1970s when they were then aged 45 to 64 years, Hart, Hole, and Smith (2000) failed to find a relationship between cholesterol and stroke incidence. This was true for both men and women considering a follow-up period of up to 20 years. Dyker, Weir, and Lees (1997) point out that some studies find that cholesterol has a positive association with ischaemic stroke and a negative association with hemorrhagic stroke, and that it is likely that when stroke is studied without subdividing the subtypes, the influence of cholesterol is reduced.

11. MORTALITY

To complete our model we need to specify the force of mortality, λ^D, for all those people who have not yet suffered a CHD event or a stroke. We assume this force of mortality depends on attained age and sex but not on any of the other risk factors. So, for example, any excess mortality appropriate for a life with diabetes is assumed to be due wholly to the increased possibility of a CHD event or stroke. We calculated this force, or intensity, as follows:

Table 7
Coefficients of the Linear Predictor for Incidence of Category 3 Hypertension

Variable	Coefficient	Value	Std. Error
	Intercept(α_{int})	-4.071	2.717×10^{-1}
	Age(β)	1.539×10^{-2}	4.198×10^{-3}
Sex			
	Males(γ_s)	-8.670×10^{-2}	3.638×10^{-2}
	Females	$-\gamma_s$	

Table 8
Coefficients of the Linear Predictor for Incidence of CHD for Males

Variable	Coefficient	Value	Std. Error
	Intercept(α_{int})	-11.75	1.485
	Age(β_0)	1.848×10^{-1}	4.921×10^{-2}
	Age^2(β_1)	-1.113×10^{-3}	4.028×10^{-4}
Hypertension			
	Categories 0 or 1(δ_0)	-5.211×10^{-1}	9.160×10^{-2}
	Category 2(δ_1)	5.935×10^{-2}	7.603×10^{-2}
	Category 3	$-(\delta_0 + \delta_1)$	
Smoking			
	No(ρ_k)	-1.317×10^{-1}	5.146×10^{-2}
	Yes	$-\rho_k$	
Hypercholesterolemia			
	Categories 0 or 1(η_{c^*})	-2.727×10^{-1}	6.098×10^{-2}
	Category 2	$-\eta_{c^*}$	
Diabetes			
	No(ϕ_d)	-1.333×10^{-1}	6.210×10^{-2}
	Yes	$-\phi_d$	

$$\lambda_x^D = \mu_x^{ELT15} \times (1 - f^{CHD}(x) - f^{STR}(x)),$$

where μ_x^{ELT15} is the force of mortality at age x according to English Life Table Number 15, Males or Females, as appropriate, and $f^{CHD}(x)$ and $f^{STR}(x)$ are factors to adjust ELT 15 to take out deaths following a CHD event or stroke, respectively. The number of deaths by cause in England and Wales in 1990, 1991, and 1992 is given for five-year age groups in OPCS (1991, 1993a, 1993b), but numbers of deaths due to all causes are available for single ages. We have used these data to estimate $f(\cdot)$ for five-year age groups. For a five-year age group centered on age x, $f^{CHD}(x)$ was estimated as

$$\frac{\text{Number of deaths in 1990, 1991, and 1992 in the age group due to CHD}}{\text{Total number of deaths in 1990, 1991, and 1992 in the age group}}.$$

These estimates were then smoothed using least squares over the age range 0–92. The factor $f^{STR}(x)$ was estimated similarly (although the age range for smoothing was 22–92).

Table 9
Coefficients of the Linear Predictor for Incidence of CHD for Females

Variable	Coefficient	Value	Std. Error
	Intercept(α_{int})	-17.00	3.784
	Age(β_0)	3.003×10^{-1}	1.174×10^{-1}
	Age^2(β_1)	-1.916×10^{-3}	8.997×10^{-4}
Hypertension			
	Categories 0 or 1(δ_0)	-8.145×10^{-1}	1.832×10^{-1}
	Category 2(δ_1)	5.794×10^{-2}	1.382×10^{-1}
	Category 3	$-(\delta_0 + \delta_1)$	
Smoking			
	No(ρ_k)	-3.195×10^{-1}	8.265×10^{-2}
	Yes	$-\rho_k$	
Hypercholesterolemia			
	Categories 0 or 1(η_{c^*})	-2.513×10^{-1}	1.183×10^{-1}
	Category 2	$-\eta_{c^*}$	
Diabetes			
	No(ϕ_d)	-2.862×10^{-1}	9.081×10^{-2}
	Yes	$-\phi_d$	

Table 10
Coefficients of the Linear Predictor for Incidence of Stroke

Variable	Coefficient	Value	Std. Error
	Intercept(α_{int})	-10.47	4.717×10^{-1}
	Age(β)	7.716×10^{-2}	7.011×10^{-3}
Hypertension			
	Categories 0, 1, or 2(δ_{b*})	-6.416×10^{-1}	6.700×10^{-2}
	Category 3	$-\delta_{b*}$	
Smoking			
	No(ρ_k)	-1.911×10^{-1}	6.293×10^{-2}
	Yes	$-\rho_k$	
Diabetes			
	No(ϕ_d)	-1.986×10^{-1}	6.809×10^{-2}
	Yes	$-\phi_d$	
Sex			
	Male(γ_s)	-7.824×10^{-1}	4.371×10^{-1}
	Female	$-\gamma_s$	
Age*Sex			
	Age:Male(ψ)	1.365×10^{-2}	6.480×10^{-3}
	Age:Female	$-\psi$	

For males, the (smoothed) functions $f^{CHD}(x)$ and $f^{STR}(x)$ are given by

$$f^{CHD}(x) = \begin{cases} \exp(-9.414 + 0.2008 \times x) & : x \le 32.5 \\ -1.479 + 0.0740 \times x - 9.478 \times 10^{-4} \times x^2 + 3.734 \times 10^{-6} \times x^3 & : x \ge 38 \end{cases}$$

with linear blending for $32.5 < x < 38$:

$$f^{STR}(x) = 0.2274 - 0.03079 \times x + 1.555 \times 10^{-3} \times x^2 - 3.478 \times 10^{-5} \times x^3 + 3.602 \times 10^{-7} \times x^4 - 1.392 \times 10^{-9} \times x^5$$

for $x \ge 20$, and 0 otherwise.

The adjustment factors for females are given by

$$f^{CHD}(x) = \exp(-9.201 + 0.2057 \times x - 1.337 \times 10^{-3} \times x^2),$$

$$f^{STR}(x) = 0.3306 - 0.04385 \times x + 2.310 \times 10^{-3} \times x^2 - 5.439 \times 10^{-5} \times x^3 + 5.878 \times 10^{-7} \times x^4 - 2.341 \times 10^{-9} \times x^5$$

for $x \ge 20$, and 0 otherwise.

Sample values of the force of mortality for ELT 15 Males and ELT 15 Females, the adjustment factor $f^{CHD}(x) + f^{STR}(x)$ and λ_x^D are given in Table 11.

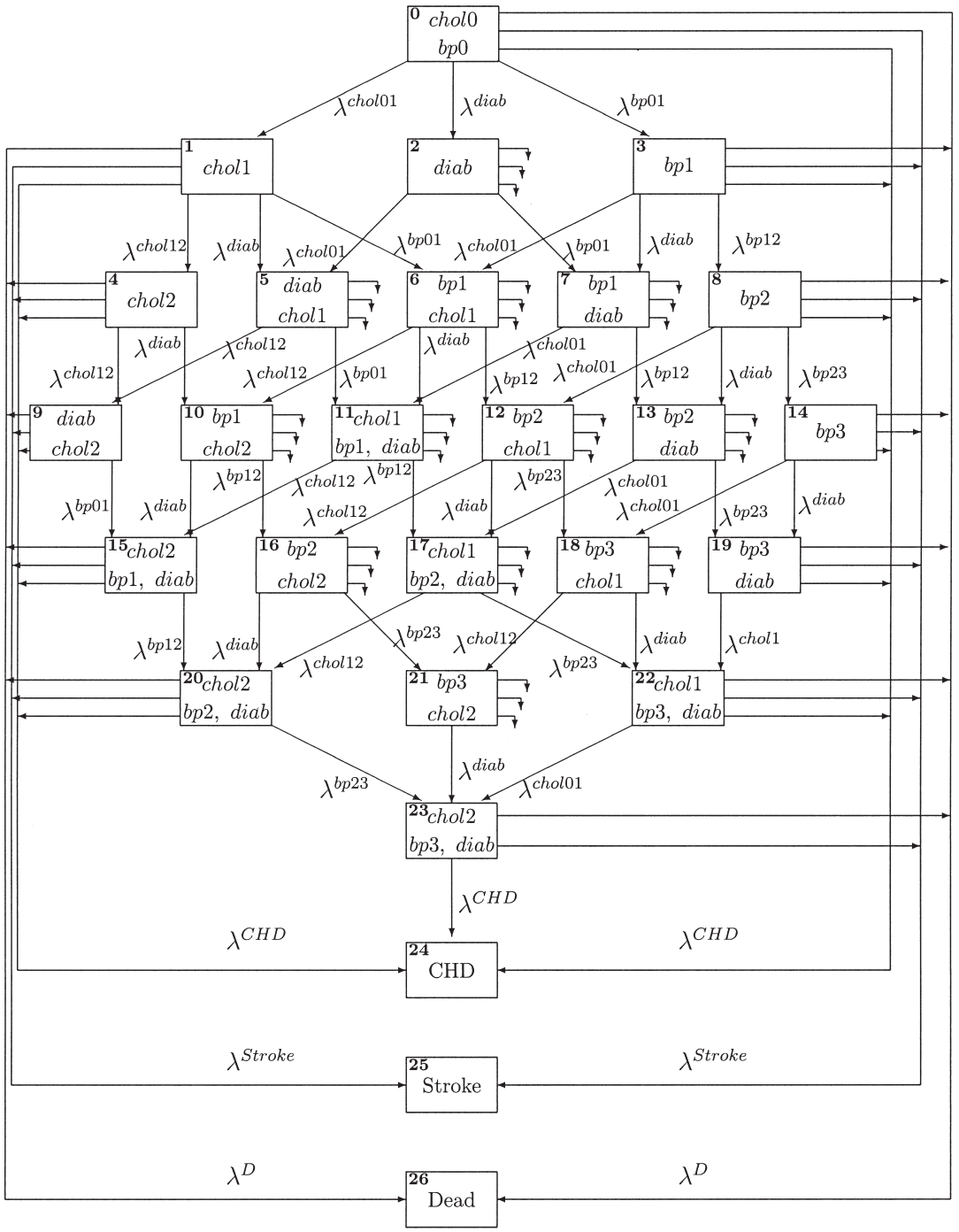
Table 11
Force of Mortality

Age	Males			Females		
	μ^{ELT15}	$f^{CHD} + f^{STR}$	λ_x^D	μ^{ELT15}	$f^{CHD} + f^{STR}$	λ_x^D
20	0.00083	0.01306	0.00082	0.00032	0.03265	0.00031
30	0.00090	0.05579	0.00085	0.00042	0.05931	0.00040
40	0.00166	0.24095	0.00126	0.00102	0.10119	0.00092
50	0.00440	0.36240	0.00281	0.00280	0.16104	0.00235
60	0.01323	0.40676	0.00785	0.00786	0.25272	0.00587
70	0.03833	0.40847	0.02267	0.02123	0.36128	0.01356
80	0.09675	0.38821	0.05919	0.05827	0.43626	0.03285

12. THE MODEL

Our Markov model for the development of CHD and stroke is shown as Figure 3. As explained earlier, we have separate parameterizations for the 12 subpopulations determined by sex (2), smoking status (2), and BMI (3). The factors retained as significant in the parameterization of the intensities are

Figure 3
CHD and Stroke Model



summarized in Table 12. Note that, for the reasons explained in Section 6, there are no "backward" transitions between the 23 transient states of the model.

Waters and Wilkie (1987) derive algorithms that can be used to calculate the probability that a life starting at a given age in a given state will be in any specified state at a specified future age. Since our model is Markov, this is equivalent to solving Kolmogorov's equations. Norberg (1995) derives a series of simultaneous differential equations for the present values of the moments of continuous payments made while in any of the states of such models and lump-sum payments on transition between any of the states. He also shows how to solve this system of equations numerically. For the first moment, this amounts to solving Thiele's equations. Norberg's algorithm will be used in Part II to calculate premium rates. In all our numerical work using these algorithms, we will use a step size of 0.01 years.

Our model is a discrete–state space, continuous–time Markov model. Such models have been shown to be extremely versatile tools for the modeling of life histories in the context of insurance and elsewhere. See, for example, Hoem (1988) and Macdonald, Waters, and Wekwete (2003a, 2003b). Advantages of these models are that parameter estimation is relatively straightforward and that numerical algorithms have been developed to calculate moments of the present values of payment streams resulting from transitions between states and sojourns in states (Norberg 1995). Similar, but not identical, models have been advocated by Tolley and Manton (1991). See also Manton et al. (1994). Interestingly, these papers used data from the Framingham study to illustrate their models.

13. COMPARISON OF MODELED AND ESTIMATED PROBABILITIES

The model developed in the earlier sections has been parameterized largely using the Framingham data. Before testing our model against more recent, and more relevant, data sets, it is of some interest to see how well it models the Framingham data itself. In particular, since our parameterization has been based on *transition intensities,* it is of interest to see how well our model reproduces the *probabilities* of CHD or stroke, the major end points for this part of our study, over a specified period of time for lives with a given risk profile at the start of the period.

Tables 13 and 14 show the probability of CHD (before a stroke) and stroke (before CHD), respectively, within 10 years for lives aged 45, 55, and 65, estimated directly from the Framingham data and using our model. We estimated these probabilities separately for each sex and smoking status. We have not distinguished between different BMI levels since BMI has relatively little impact on these probabilities—the probabilities calculated using the model have assumed normal BMI. The probabilities are also shown separately for each risk profile, indicated by which of the 23 transient states of our model these individuals are in at the start of the 10-year period. The probabilities estimated from the Framingham data are based on aggregating lives into 10-year age groups centered on the three starting ages, 45, 55, and 65, and estimates have been included in the tables only if there were at least five cases of CHD or

Table 12
Summary of Models

Function	Age	Sex	Smoking	BMI	Blood Pressure	Cholesterol	Diabetes
λ^{CHD}	•	•	•		•	•	•
λ^{Stroke}	•	•	•		•		•
λ^{bp01}	•			•			
λ^{bp12}	•	•					
λ^{bp23}	•	•					
λ^{chol01}_{males}							
$\lambda^{chol01}_{females}$	•						
λ^{chol12}	•	•					
λ^{diab}	•			•			

Table 13
Estimated and Modeled 10-Year Probabilities of CHD

Starting State	Age	LCL	Estimated Probability	UCL	Modeled Probability
Males					
Nonsmokers					
12	55	0.024	0.133	0.242	0.061
21	55	0.043	0.109	0.175	0.127
23	55	0.081	0.276	0.471	0.159
10	65	0.013	0.119	0.226	0.092
16	65	0.063	0.148	0.234	0.135
21	65	0.117	0.195	0.273	0.176
23	65	0.138	0.309	0.481	0.214
Smokers					
1	45	0.004	0.041	0.078	0.028
10	45	0.052	0.116	0.180	**0.050**
12	45	0.005	0.052	0.098	0.051
21	45	0.063	0.148	0.234	0.107
4	55	0.027	0.082	0.136	0.066
10	55	0.028	0.083	0.138	0.078
12	55	0.049	0.124	0.198	0.076
16	55	0.057	0.106	0.154	0.117
20	55	0.020	0.185	0.351	0.148
21	55	0.098	0.155	0.211	0.157
10	65	0.023	0.128	0.232	0.114
12	65	0.027	0.111	0.195	0.106
16	65	0.051	0.120	0.189	0.166
18	65	0.093	0.234	0.375	0.138
21	65	0.076	0.141	0.205	**0.213**
23	65	0.049	0.200	0.351	0.256
Females					
Nonsmokers					
21	55	0.008	0.028	0.049	0.041
16	65	0.019	0.044	0.070	0.041
21	65	0.042	0.069	0.096	0.065
23	65	0.055	0.149	0.244	0.103
Smokers					
16	55	0.032	0.076	0.121	0.047
21	55	0.012	0.051	0.089	0.076
21	65	0.046	0.099	0.152	0.117

stroke, as appropriate. Also shown in these tables are the lower (LCL) and upper (UCL) 95% confidence limits for the estimates based on the data.

Values of the probabilities calculated from the model are shown in bold type where they lie outside the 95% confidence limits. It can be seen that we have only two such cases out of 30 probabilities for CHD and two cases out of 13 probabilities for stroke. Tables showing 5- and 15-year probabilities of CHD and stroke are given by Wekwete (2002, Tables 4.98, 4.100, 4.101, and 4.103). These show an even better fit of our model to the data than is shown in Tables 13 and 14: only one out of a total of 78 probabilities lies outside the 95% confidence limits.

14. Comparisons with Other Data Sets

In this section we use some data sets more recent than the Framingham data, and that relate to U.K. experience, to derive rates that can then be compared with rates derived from the Framingham data or with rates calculated from our model. These other data sets could not be used to parameterize our model since they are far less comprehensive than the Framingham data. Our purpose in this section is to determine where, if at all, our parameterization needs to be adjusted for our model to be applicable in

Table 14
Estimated and Modeled 10-Year Probabilities of Stroke

Starting State	Age	LCL	Estimated Probability	UCL	Modeled Probability
Males					
Nonsmokers					
16	65	0.007	0.062	0.117	0.036
21	65	0.004	0.039	0.074	**0.079**
Smokers					
21	55	0.030	0.067	0.104	0.052
16	65	0.023	0.080	0.137	0.050
21	65	0.059	0.119	0.178	0.110
Females					
Nonsmokers					
12	55	0.006	0.056	0.106	0.016
21	55	0.002	0.018	0.034	**0.035**
18	65	0.027	0.093	0.159	0.064
23	65	0.016	0.090	0.163	0.086
Smokers					
21	55	0.012	0.051	0.089	0.050
16	65	0.011	0.062	0.112	0.042
21	65	0.017	0.057	0.097	0.088
23	65	0.023	0.217	0.412	0.117

the United Kingdom at the present time. Other authors have studied the applicability of the Framingham data to other populations. Haq et al. (1999) conclude that "the Framingham (risk) function separates high and low CHD risk groups and is acceptably accurate for northern European populations, at least in men."

14.1 Morbidity Statistics from General Practice

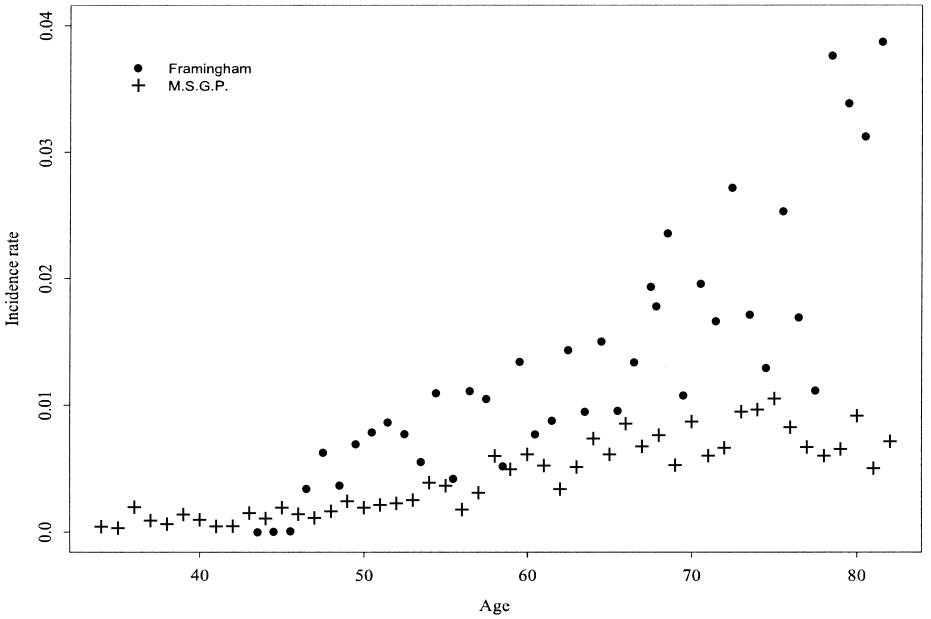
The Morbidity Statistics from General Practice (MSGP) fourth national study (McCormick, Fleming, and Charlton 1995) was carried out from 1 September 1991 to 31 August 1992 with 60 practices in England and Wales participating. The survey covered 502,000 people whose characteristics were representative of the population of England and Wales as given by the 1991 Census and the General Household Survey. The survey recorded cases of first ever consultation for illnesses as well as the age, sex, and social class, among other characteristics, of the registered patients. For most patients above 16 years of age, the survey also recorded their smoking status. However, the patient's BMI was not recorded. A problem with the MSGP data is that diagnoses of, for example, hypertension depend on the patient deciding to visit their general practitioner and then on the latter's subjective judgment.

14.1.1 Diabetes

Figure 4 shows incidence rates of diabetes derived from the Framingham data and from the MSGP data. For the former, the threshold for diabetes was taken to be a blood sugar level of 140 mg/dL. We considered this to be more likely to be consistent with the definition of diabetes commonly used in 1991–92 (exact blood sugar levels were not recorded by McCormick, Fleming, and Charlton 1995). The rates are shown by age but have been aggregated over all other factors, for example, sex, smoking status, and BMI. (Recall from Table 12 that only age and BMI were significant factors for the incidence of diabetes.)

Figure 4 shows that the incidence rates of diabetes from the Framingham data are consistently higher than those from the MSGP data. In part, this difference may be due to an underreporting of diabetes, a feature that is likely to affect the latter rates but not the former, for the following reasons:

Figure 4
Incidence Rates for Diabetes Derived from Framingham Data and Morbidity Statistics from General Practice



1. The initial symptoms of diabetes may not be severe, so the condition may go undetected for some time.
2. To be recorded in the MSGP data, patients have to refer themselves to their general practitioner.

It has been estimated that between one-third and one-half of all diabetes cases are undiagnosed at any given time (see Harris et al. 1998; Lawrence et al. 2001). This underreporting may also explain the shape of the incidence curve for the MSGP data, which falls at the older ages.

Doubling the MSGP incidence rates would, very roughly, bring them into line with the Framingham rates over the range of ages of most interest to us in terms of CI insurance, say, ages 30–70. Given the comments in the previous paragraph about the underreporting of diabetes in the MSGP data, our conclusion is that the parameterization of the intensity of developing diabetes set out in Section 8.1 is not unreasonable for the application of our model to CI insurance in the United Kingdom at the present time.

14.1.2 Hypertension

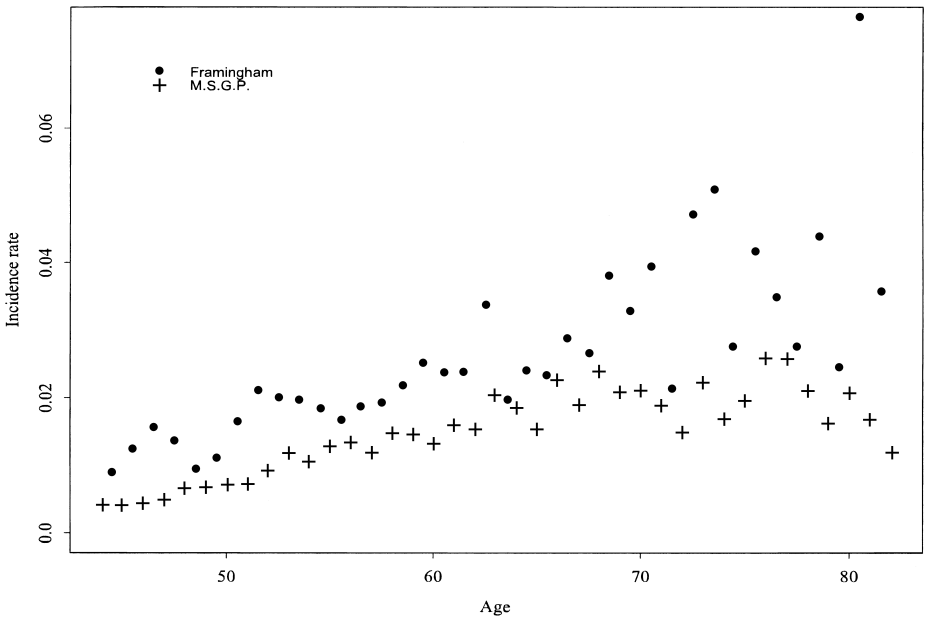
Figure 5 shows incidence rates of hypertension derived from the Framingham data and from the MSGP data. For the former, the threshold for hypertension was taken to be systolic blood pressure of 160 mmHg. As was the case for Figure 4, the rates are shown by age but have been aggregated over all other factors.

The Framingham rates are generally higher than those from the MSGP data. This may be due in part to precisely the same factors as those listed in Section 14.1.1 in relation to diabetes.

14.2 The Health Survey for England

The Health Survey for England is an annual exercise that concentrates on some aspect of the health of people in England. The 1998 Survey (Erens and Primatesta 1999) concentrated on cardiovascular disease. It is based on a sample designed to be representative of the population of England aged two or more who live in private households.

Figure 5
Incidence Rates for Hypertension Derived from Framingham Data and Morbidity Statistics from General Practice



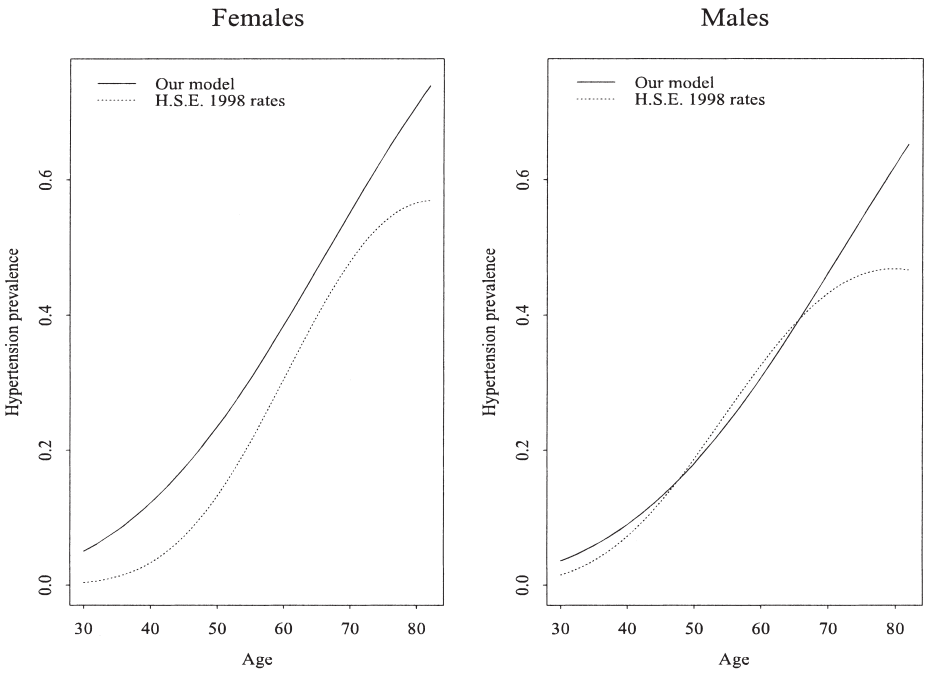
14.2.1 Hypertension

Figure 6 shows, separately for females and males, prevalence rates of hypertension based on the Health Survey for England 1998 (HSE) (Erens and Primatesta 1999) and our model. The definition of hypertension for the HSE rates is systolic blood pressure in excess of 160 mmHg or diastolic blood pressure in excess of 95 mmHg or being treated for high blood pressure. We regard this as broadly equivalent to our Category 3 for hypertension.

The prevalence rates based on our model have been calculated as follows. We constructed a model with five states: Categories 0, 1, 2, and 3 for hypertension and "Dead." The transition intensity between Categories i and j for hypertension, where $0 \leq i < j \leq 3$, is λ_x^{bpij}, as described in Section 8.3. Note that, as in the model in Figure 3, no "backward" transitions are possible between these categories. From each category for hypertension it is possible to die. The force of mortality is that of ELT 15, Males or Females, as appropriate, from each of the four categories for hypertension. Assuming a life aged 0 started in Category 0 for hypertension, we calculated the probability that at any future age the life was in the state corresponding to Category 3 for hypertension, given that the life was still alive. The resulting age specific prevalence rates depend on sex and BMI (note that transitions into and between categories of hypertension do not depend on smoking status, as shown in Table 12). We produced prevalence rates for each sex as weighted averages of the rates for the three BMI categories using weights given by Erens and Primatesta (1999); see Table 15.

The procedure outlined in the previous paragraph could be criticized on the grounds that the force of mortality should depend on the category for hypertension: that is, it should be higher for the higher categories, other factors being equal. We recalculated the age-specific prevalence rates for hypertension assuming that the force of mortality for individuals in the state corresponding to Category 3 for hypertension was twice, and then three times, the level of ELT 15. This made very little difference to the prevalence rates below age 65, where mortality rates are relatively low. However, above age 65 the prevalence rates were reduced, significantly at very high ages, compared to those in Figure 6.

Figure 6
Comparison of Hypertension Prevalence Rates



Our conclusion from the comments in Section 14.1.2 and from Figure 6 is that our parameterization of the incidence of Categories 1, 2, and 3 for hypertension for males is not unreasonable in terms of the applications of our model in Part II, but may be a little high for females.

14.2.2 Hypercholesterolemia

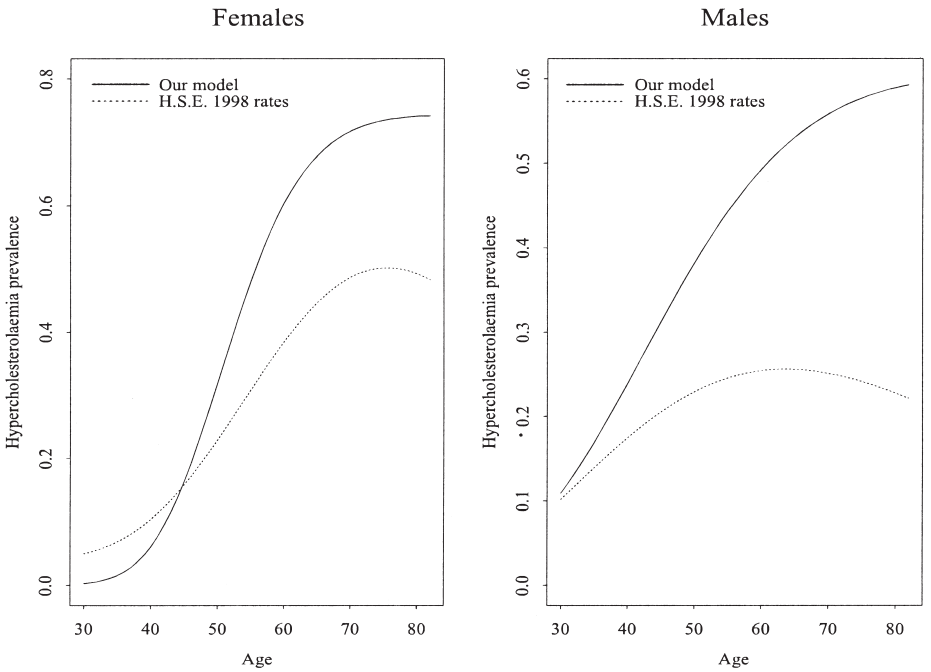
Figure 7 shows separately for females and males age-specific prevalence rates of hypercholesterolemia based on the Health Survey for England (Erens and Primatesta 1999) and based on our model. Erens and Primatesta (1999) define hypercholesterolemia as total cholesterol exceeding 251 mg/dL. The prevalence rates derived from our model have been calculated in the same way as the prevalence rates for hypertension (see Section 14.2.1), with "hypercholesterolemia" taken to be Category 2 for hypercholesterolemia, so that our threshold for hypercholesterolemia is 240 mg/dL.

It can be seen from Figure 7 that the prevalence rates for hypercholesterolemia derived from our model are much higher than those given by Erens and Primatesta (1999), particularly for males at higher ages, although the difference in thresholds is a factor here. Table 16, taken from Erens and Primatesta (1999), is interesting because it shows the change in the age/sex-specific prevalence rates for hypercholesterolemia in England between 1994 and 1998. The fall in these *prevalence* rates in just four

Table 15
Subdivision (Proportions) of Population of England by BMI Category

	Males	Females
BMI $\leq$ 25	0.37	0.47
25 $<$ BMI $\leq$ 30	0.46	0.32
BMI $>$ 30	0.17	0.21

Figure 7
Comparison of Hypercholesterolemia Prevalence Rates



years is remarkable. At this stage it is worth recalling the discussion in Section 8.2 on the incidence rates for hypercholesterolemia in the Framingham data. Health, United States (2003, Table 67) shows the mean serum cholesterol level in a sample of the population of the United States aged 20–74 at roughly 10-year intervals over the period 1960–2000. There is a steady downward trend for both sexes starting from 220 mg/dL for males and 224 mg/dL for females in 1960 and reducing by around 10% to 204 mg/dL for males and 203 mg/dL for females by 2000.

We recalculated the age-specific prevalence rates for hypercholesterolemia assuming the force of mortality from the state corresponding to Category 2 for hypercholesterolemia is 1.5 times, and then twice, that of ELT 15. The effects were the same as for the prevalence of hypertension: very little change in the prevalence rates below age 65 compared to Figure 7, but a drop in the prevalence rates above age 65.

It seems clear that for the applications of this model in Part II, we should reduce the levels of the intensities of Categories 1 and/or 2 for hypercholesterolemia, as set out in Section 8.2, for both males and females.

Table 16
Prevalence of Hypercholesterolemia Based on Health Survey for England

	Age Group							
	16–24	**25–34**	**35–44**	**45–54**	**55–64**	**65–74**	**75+**	**All Ages**
Males (1998)	0.019	0.108	0.169	0.238	0.229	0.264	0.202	0.18
Females (1998)	0.029	0.067	0.088	0.220	0.374	0.480	0.444	0.224
Males (1994)	0.036	0.147	0.309	0.393	0.407	0.383	0.301	0.279
Females (1994)	0.048	0.098	0.134	0.322	0.574	0.674	0.576	0.319

14.3 Hospital Episodes Statistics 1993–94

Dinani et al. (2000) use the Hospital Episodes Statistics 1993–94 data to calculate incidence rates for CHD events, which they define as acute myocardial infarction, and stroke, which they define as cerebrovascular disease, specifically excluding transient ischaemic attacks. These rates are adjusted so that they relate to "new and first ever" events, and the stroke rates are adjusted so that they exclude people who have already experienced a CHD event. These rates are age and sex specific, but are not subdivided by any other characteristics.

For each of the 12 subpopulations defined by sex, smoking status, and BMI, we calculated using the model in Figure 3 the probability of being in each of the states 0–23 at each future age, given that the life started in state 0 at age 30 and given that the life had not yet died, had a CHD event or had a stroke. We used these occupancy probabilities to weight the intensities of a CHD event, λ^{CHD}, and stroke, λ^{Stroke}, to obtain age-specific weighted averages of these intensities. These intensities were then averaged over the six combinations of smoking status and BMI. The proportions in each BMI category are as shown in Table 15, and it was assumed that 27% of females and 28% of males at all ages are smokers, independent of BMI category (see Erens and Primatesta 1999). The resulting age- and sex-specific incidence rates are compared with those from Dinani et al. (2000, Tables 3.3(m), 3.3(f), 3.4(m), and 3.4(f), "Smoothed, Adjusted Crude Rates") in Figure 8.

15. Adjustments to the Intensities of Hypertension and Hypercholesterolemia

In this section we investigate the effect of reducing the intensities of hypertension (females) and hypercholesterolemia (both sexes) on the prevalence rates for these two conditions and on the incidence rates of CHD and stroke.

Figures 9 and 10 show the prevalence rates for hypertension and for hypercholesterolemia with the following adjustments applied to the parameterizations of the intensities given in Section 8:

Figure 8
Comparison of CHD and Stroke Incidence Rates from Our Model and Those from Dinani et al. (2000)

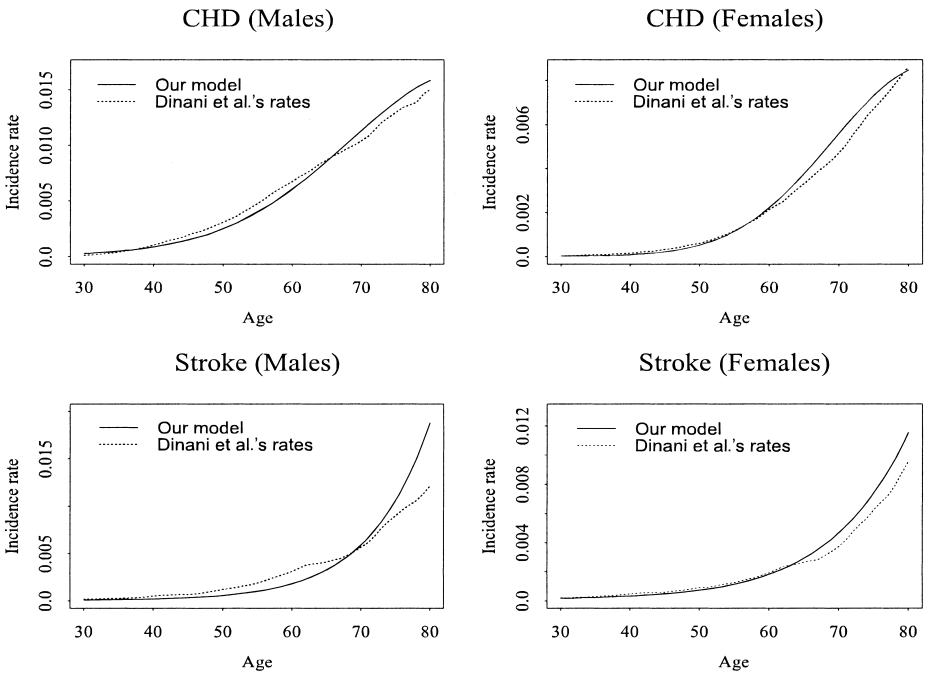


Figure 9
Comparison of Hypertension Prevalence Rates: Adjusted Intensities

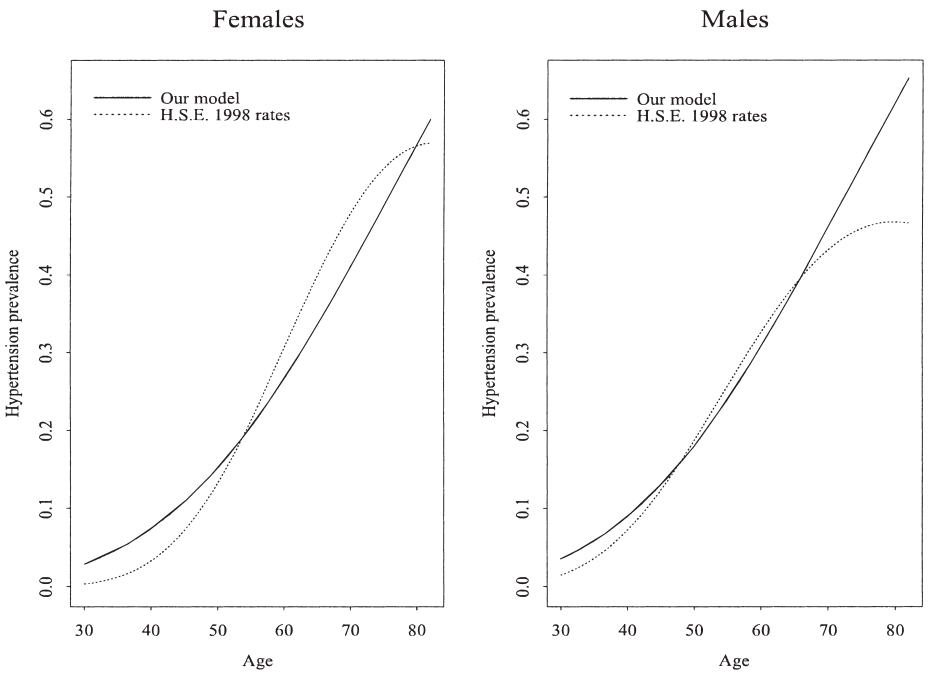
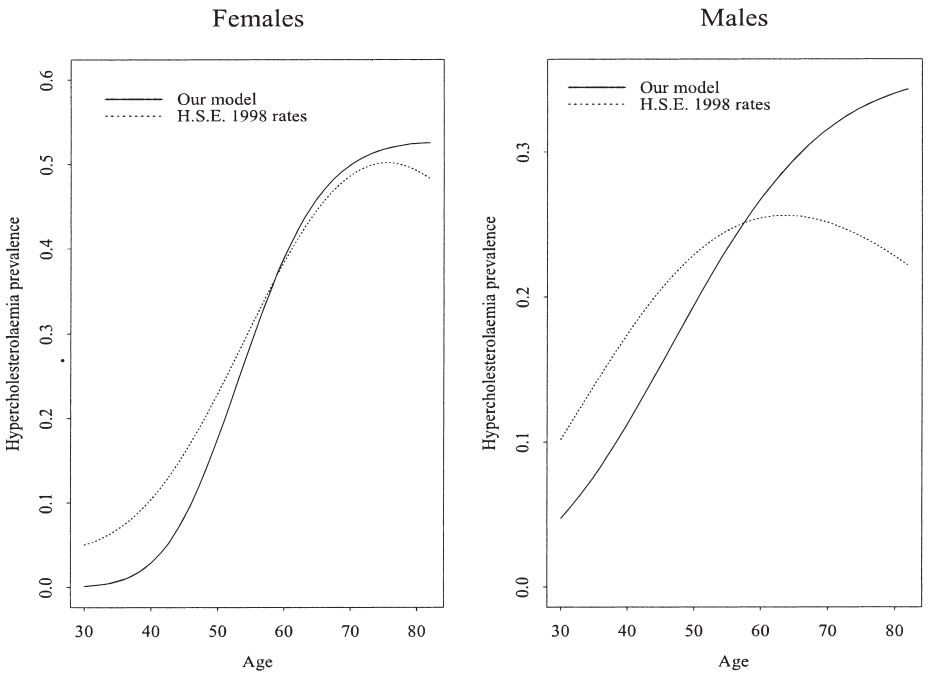


Figure 10
Comparison of Hypercholesterolemia Prevalence Rates: Adjusted Intensities



1. For females, the intensities for all categories of hypertension, λ^{sbp01}, λ^{sbp12}, and λ^{sbp23}, have been multiplied by 0.8. The rates for males are unchanged.
2. The intensities for all categories of hypercholesterolemia, λ^{chol01} and λ^{chol12}, have been multiplied by 0.65 (females) and 0.6 (males).

Figures 9 and 10 should be compared with Figures 6 and 7. It can be seen that the former, incorporating the adjusted intensities, give a better fit to the rates taken from Erens and Primatesta (1999), that is, the rates from the Health Survey for England. The prevalence rates for hypercholesterolemia are still a little higher than the Health Survey for England rates over the range of ages of interest to us, but the difference in threshold values for the two sets of rates needs to be borne in mind.

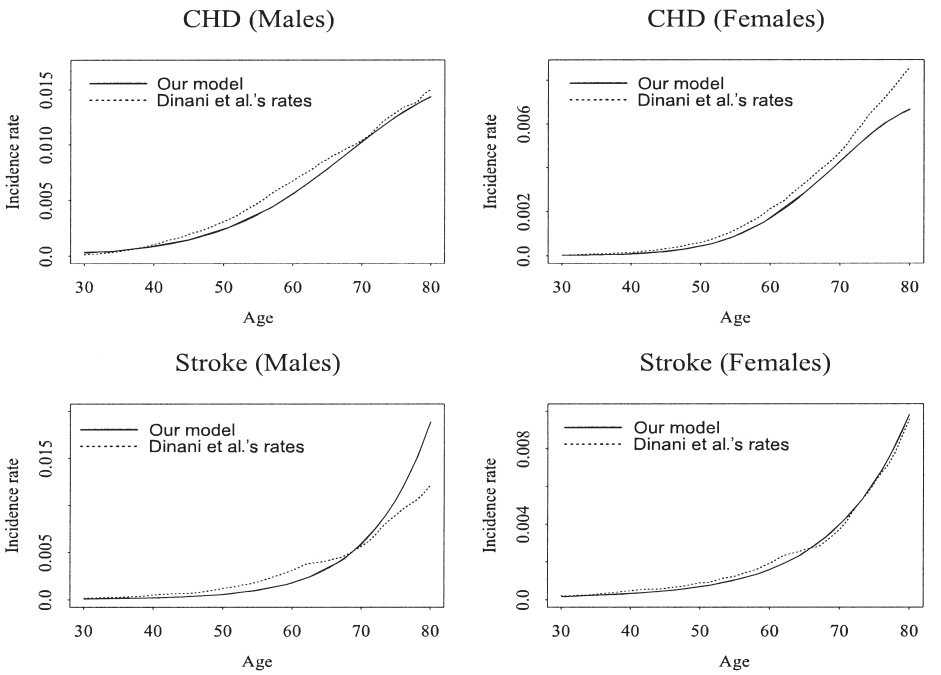
Figure 11 compares the (weighted average) incidence rates for CHD and stroke, calculated using the adjusted intensities, with those in Dinani et al. (2000). This figure should be compared with Figure 8. The adjustments generally tend to decrease the incidence of CHD and stroke, but it can be seen the differences are marginal.

16. OTHER MODELS

The model developed in this part is a model for the development of CHD and stroke, incorporating known risk factors. Similar models have been developed within the medical/health management sciences. A good example of such a model is described by Babad et al. (2002). They use the (original cohort) Framingham data as a basis for the parameterization of their model and then use more recent U.K. data to "calibrate the baseline risk to the English level." In this respect their approach is identical to ours. The differences between their study and ours are the following:

a. They develop a discrete-time model, whereas ours operates in continuous time.

Figure 11
Comparison of CHD and Stroke Incidence Rates from Our Model and Those from Dinani et al. (2000): Adjusted Intensities



b. They are concerned with intervention strategies to reduce the risk of CHD, whereas our motivation is the quantification of insurance risk.
c. A consequence of the previous point is that Babad et al. (2002) consider a range of CHD outcomes, for example, stable angina, which are not included in our model because they would not trigger a CI insurance claim, but they do not include stroke in their model.

The above comments also apply to an earlier model developed by Weinstein et al. (1987), which, additionally, did not include diabetes as a risk factor for CHD.

17. CONCLUSIONS

In this part we have described and parameterized a model for the development of CHD or stroke. Our applications of this model in Part II will be to the underwriting of CI insurance, and we have developed our model with these applications in mind. For example, we have included in our model only those risk factors for CHD and/or stroke typically taken into account by CI underwriters. Our model is in some respects similar to models that have been developed for health care management and could be extended, for example, by including more CHD conditions such as stable angina. The major factor affecting the development of the model is the paucity of good quality data for the parameterization of the intensities of moving between states.

ACKNOWLEDGMENTS

This work was funded by Swiss Re Life and Health, to whom we are grateful for financial support and for many discussions with their actuarial, medical, and underwriting staff and consultants. In particular, we wish to thank Douglas Keir, Dr. David Muiry, Dr. Kevin Somerville, and Dr. Hanspeter Würmli.

The authors are grateful to two anonymous referees whose constructive comments helped to improve this paper.

This paper, and Part II, use data supplied by the National Heart, Lung, and Blood Institute, NIH, DHHS. The views expressed in these papers are those of the authors and do not necessarily reflect the views of the National Heart, Lung, and Blood Institute. We also use data supplied by the United States Renal Data Systems (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government.

REFERENCES

American Diabetes Association. 2000. Clinical Practice Recommendations 2000: Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* 22, Supplement 1.

Babab, Hannah, Colin Sanderson, Bhash Naidoo, Ian White, and Duolao Wang. 2002. The Development of a Simulation Model of Primary Prevention Strategies for Coronary Heart Disease. *Health Care Management Science* 5: 269–74.

Bamford, John, Peter Sandercock, Martin Dennis, Charles Warlow, Lesley Jones, Klim McPherson, Martin Vessey, Godfrey Fowler, Andrew Molyneux, Trevor Hughes, John Burn, and Derick Wade. 1988. A Prospective Study of Acute Cerebrovascular Disease in the Community: The Oxfordshire Community Stroke Project 1981–86. 1: Methodology, Demography and Incident Cases of First-Ever Stroke. *Journal of Neurology, Neurosurgery and Psychiatry* 51: 1373–80.

Brackenridge, R. Douglas, and W. John Elder. 1998. *Medical Selection of Life Risks.* 4th ed. New York: Macmillan Reference.

Dinani, Azim, David Grimshaw, Neil Robjohns, Stephen Somerville, Alasdair Spry, and Jerry Staffurth. 2000. *A Critical Review: Report of the Critical Illness Healthcare Study Group.* Unpublished manuscript presented to the Staple Inn Actuarial Society, London, 14 March 2000.

Dyker, Alexander G., Christopher J. Weir, and Kennedy R. Lees. 1997. Influence of Cholesterol on Survival after Stroke: A Retrospective Study. *British Medical Journal* 314: 1584–88.

Erens, Bob, and Paola Primatesta, eds. 1999. *Health Survey for England. Cardiovascular Disease 1998.* Volume 1. London: Stationery Office.

Gubitz, Gord, and Peter Sandercock. 2000. Extracts from 'Clinical Evidence': Acute Ischaemic Stroke. *British Medical Journal* 320: 692–96.

Haq, Ihtshamul U., Lawrence E. Ramsay, Wilfrid W. Yeo, Peter R. Jackson, and Erica J. Wallis. 1999. Is the Framingham Risk Function Valid for Northern European Populations? A

Comparison of Methods for Estimating Absolute Coronary Risk in High Risk Men. *Heart* 81: 40–46.

HARRIS, MAUREEN I., KATHERINE M. FLEGAL, CATHERINE C. COWIE, MARK S. EBERHART, DAVID E. GOLDSTEIN, RANDIE R. LITTLE, HSIO-MEY WIEDMEYER, AND DANITA D. BYRD-HOLT. 1998. Prevalence of Diabetes, Impaired Fasting Glucose and Impaired Glucose Tolerance in U.S. Adults. The Third National Health and Nutrition Examination Survey, 1988–1994. *Diabetes Care* 21: 518–24.

HART, CAROLE L., DAVID HOLE, AND GEORGE DAVEY SMITH. 2000. Comparison of Risk Factors for Stroke Incidence and Mortality in 20 Years of Follow-up in Men and Women in the Renfrew/Paisley Study in Scotland. *Stroke* 31: 1893–96.

HEALTH, UNITED STATES. 2003. Washington, D.C.: National Center for Health Statistics. Online at www.cdc.gov/nchs/data/hus/hus03.pdf.

HILL, TONY, AND JULIAN ROBERTS. 1996. Changing the Threshold of Body Mass Index That Indicates Obesity Affects Health Targets. *British Medical Journal* 313: 815–16.

HOEM, JAN. 1988. The Versatility of the Markov Chain as a Tool in the Mathematics of Life Insurance. *Transactions of the 1988 International Congress of Actuaries* R, 171–202.

KARDIA, SHARON L. R., MARTHA B. HAVILAND, ROBERT E. FERRELL, AND CHARLES F. SING. 1999. The Relationship between Risk Factor Levels and Presence of Coronary Artery Calcification Is Dependent on Apoliprotein E Genotype. *Arteriosclerosis, Thrombosis and Vascular Biology* 19: 427–35.

KING, RICHARD A., JEROME I. ROTTER, AND ARNO G. MOTULSKY. 1992. *The Genetic Basis of Common Diseases.* Oxford Monographs on Medical Genetics No. 20. Oxford: Oxford University Press.

LAWRENCE, JAMES M., PAUL BENNETT, ALAN YOUNG, AND ANTHONY ROBINSON. 2001. Screening for Diabetes in General Practice: A Cross Sectional Population Study. *British Medical Journal* 232: 548–51.

MCCORMICK, ANNA, DOUGLAS FLEMING, AND JOHN CHARLTON. 1995. *Morbidity Statistics from General Practice. Fourth National Study 1991–1992.* Series MB5 No. 3. OPCS, Government Statistical Service.

MACDONALD, ANGUS S. 2000. Human Genetics and Insurance Issues. In *Bio-ethics for the New Millennium,* edited by I. Torrance. Edinburgh: St. Andrew Press.

———. 2001. Genetic Information and Insurance. *Genetics Law Monitor* 2(1): 1–5.

MACDONALD, ANGUS S., HOWARD R. WATERS, AND CHESSMAN T. WEKWETE. 2003a. The Genetics of Breast and Ovarian Cancer I: A Model of Family History. *Scandinavian Actuarial Journal,* 1–27.

———. 2003b. The Genetics of Breast and Ovarian Cancer II: A Model of Critical Illness Insurance. *Scandinavian Actuarial Journal,* 28–50.

MANTON, KENNETH G., ERIC STALLARD, MAX A. WOODBURY, AND JOHN E. DOWD. 1994. Time-Varying Covariates in Models of Human Mortality and Aging: Multidimensional Generalizations of the Gompertz. *Journal of Gerontology* 49: B169–90.

NATIONAL CHOLESTEROL EDUCATION PROGRAM. 2001. *Third Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation and Treatment of High Cholesterol in Adults (Adults Treatment Panel III).* National Institute of Health Publication, 0–40. Washington, D.C.: Government Printing Office.

NATIONAL HIGH BLOOD PRESSURE EDUCATION PROGRAM. 1997. Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure. *Archives of Internal Medicine* 157: 2413–46.

NORBERG, RAGNAR. 1995. Differential Equations for Moments of Present Values in Life Insurance. *Insurance: Mathematics and Economics* 17: 171–80.

NYBOE, JORGEN, GORM JENSEN, MERETE APPLEYARD, AND PETER SCHNOHR. 1991. Smoking and the Risk of First Acute Myocardial Infarction. *American Heart Journal* 122: 438–47.

OPCS. 1991. *1990 Mortality Statistics: Cause.* Series DH2 No. 17.

———. 1993a. *1991 Mortality Statistics: Cause.* Series DH2 No. 18.

———. 1993b. *1992 Mortality Statistics: Cause.* Series DH2 No. 19.

SING, CHARLES F., AND PATRICIA P. MOLL. 1990. Genetics of Atherosclerosis. *Annual Review of Genetics* 24: 171–87.

SWISS RE UNDERWRITING MANUAL. 1995. *Swiss Re Life and Health.* London.

TOLLEY, H. DENNIS, AND KENNETH G. MANTON. 1991. Intervention Effects among a Number of Risks. *Transactions of the Society of Actuaries* 43: 443–68.

WATERS, HOWARD R., AND A. DAVID WILKIE. 1987. A Short Note on the Construction of Life Tables and Multiple Decrement Tables. *Journal of the Institute of Actuaries* 114: 569–80.

WEINSTEIN, MILTON C., PAMELA G. COXSON, LAWRENCE W. WILLIAMS, THEODORE M. PASS, WILLIAM B. STASON, AND LEE GOLDMAN. 1987. Forecasting Coronary Heart Disease Incidence, Mortality and Cost: The Coronary Heart Disease Policy Model. *American Journal of Public Health* 77(11): 1417–26.

WEKWETE, CHESSMAN T. 2002. *Genetics and Critical Illness Insurance Underwriting: Models for Breast Cancer and Ovarian Cancer and for Coronary Heart Disease and Stroke.* Ph.D. thesis, Heriot-Watt University, Edinburgh. Online at www.ma.hw.ac.uk/ams.html.

WOLF, PHILIP A., RALPH B. D'AGOSTINO, WILLIAM B. KANNEL, RUTH BONITA, AND ALBERT J. BELANGER. 1988. Cigarette Smoking as a Risk Factor for Stroke: The Framingham Study. *Journal of the American Medical Association* 259: 1025–29.

Discussions on this paper can be submitted until July 1, 2005. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.