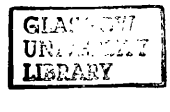
Record Linkage: Applied to a Clinical Trial and Cohort Study

Margaret Catriona Morag MacLeod

Submitted to the
University of Glasgow
for the Degree of
Doctor of Philosophy

Department of Statistics,
University of Glasgow.
December 1995

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346



Abstract

The main aim of this thesis is to investigate the role of computerised record linkage in clinical trial and epidemiological follow-up. This is illustrated using the West of Scotland Coronary Prevention Study (WOSCOPS). This study is placed in context by reviewing the results of previous clinical trials and epidemiological studies. Details of the probabilistic basis for record linkage techniques and the practical methods used to set up a computerised record linkage system are also given. The above ideas are brought together in the application of record linkage techniques, as employed in the Scottish Record Linkage System, to link the WOSCOPS subjects to their morbidity and mortality records on the Scottish national databases. The data resulting from the linkages are considered in various ways. A comparative study was carried out for the subjects randomised into WOSCOPS. The availability of two separate adverse event databases, one produced by routine subject follow-up and the other derived by computerised record linkage, provided a unique opportunity to assess the completeness and accuracy of each of the follow-up methods, and the benefits of a system incorporating both methods. This study found that record linkage compared well with traditional methods of follow-up in terms of completeness, accuracy, speed and cost. Record linkage provided the only feasible method by which adverse event records could be obtained for the large cohort of subjects screened for WOSCOPS. The data for the screened cohort were analysed in relation to categories of baseline risk factors. Data were categorised to maintain subject anonymity since informed consent was not available for all screenees. Analysis of this large cohort provided results which were in agreement with the previous studies. The mortality rates observed for the screened cohort using record linkage were finally compared to the mortality rates for the general population in the screening area. This provides an assessment of how representative the screened cohort is of the general population. It was found that the general population in the screening area tended to have higher mortality rates than the screened cohort.

Table of Contents

List of Figures

Chapter 1: Introduction

Chapter 2: The West of Scotland Coronary Prevention Study

Chapter 4: Record linkage and WOSCOPS

Chapter 5: Record linkage comparative study

Chapter 6: Analysis of linkage data for the WOSCOPS cohort

Chapter 7: Comparison of screenee and population event rates

List of Tables

Chapter 1: Introduction

Chapter 2: The West of Scotland Coronary Prevention Study

Chapter 4: Record linkage and WOSCOPS

Chapter 5 : Record linkage comparative study

Chapter 6: Analysis of linkage data for the WOSCOPS cohort

Chapter 7: Comparison of screenee and population event rates

Acknowledgements

I would like to thank my supervisor, Professor Ian Ford, for his advice and encouragement throughout the three years leading to the submission of this thesis, and for providing excellent resources. I would also like to thank him for provision of funding for my postgraduate studies.

I would also like to record my gratitude to the following people:

Dr Steve Kendrick and his staff at the Scottish Record Linkage System (especially Mr James Boyd, now at Fife Health Board); for encouragement throughout this PhD, and practical advice on setting up the linkage exercises.

Dr Andrew Whitehouse, a clinical research associate on the West of Scotland Coronary Prevention Study; for his assistance in the verification of linked records and guidance in all things medical.

All staff on the West of Scotland Coronary Prevention Study; for their efforts in the production of the datasets which were used in this project.

Mr Arthur Jones of Greater Glasgow Health Board; for his assistance in the postcoding of addresses for all the WOSCOPS subjects.

The Information and Statistics Division of the Common Services Agency of the National Health Service in Scotland; for provision of morbidity data.

The Registrar General for Scotland; for provision of mortality and census data.

Mr Mike Muirhead at the Information and Statistics Division; for provision of computer tables of disease and operation codes.

Dr Harry Burns, Director of Public Health, Greater Glasgow Health Board; for advice, and provision of a file of Carstairs Deprivation scores.

Chapter 1

Introduction

1.1 Risk factors for coronary heart disease

1.1.1 Overview

Coronary heart disease (CHD) became a disease of epidemic proportions in the West as economic development progressed and deaths due to infectious diseases decreased (Marmot, 1992). In the UK this transition occurred in the 1920s. CHD and cancer combined caused approximately the same number of deaths as infectious diseases in 1921, but by 1931 they had become the major causes of death in both men and women. However, in the West, coronary heart disease is now in decline (Figure 1.1), particularly in higher socio-economic groups, while it is on the increase in developing countries, with Eastern Europe now showing the same decline in life expectancy that Western Europe did until the 1950s and 1960s (Marmot, 1992). The observed decrease in CHD in the higher socio-economic groups in Western Europe is thought to be mainly due to decreases in dietary cholesterol and saturated fat intake, decreases in cigarette smoking, decreases in blood pressure levels due to increased use of antihypertensive drug treatment, and increased leisure time physical activity in the more educated population strata.

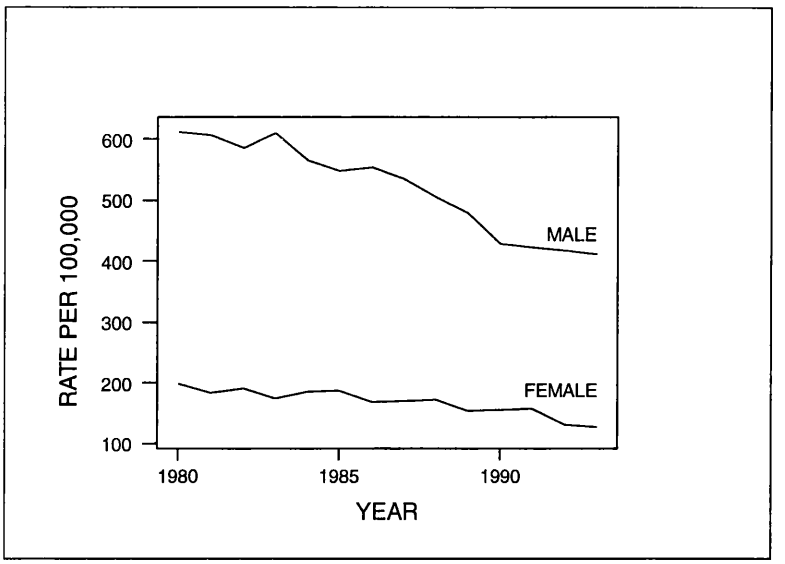


Figure 1.1 Trend in coronary heart disease deaths in Scotland for Males and Females age 45-64 from 1980 to 1993 (Data from Annual Reports of the Registrar General for Scotland).

Risk factors for coronary heart disease which are unmodifiable include age, sex and a family history of CHD. Older men with a family history of CHD are at greater risk. Diabetes is also a risk factor although it is of much lower prevalence in the population than the other risk factors. Other coronary heart disease risk factors are modifiable. These include cigarette smoking, hypertension, obesity and elevated blood cholesterol. Hypertension, obesity and serum cholesterol are all influenced by 'rich diet'. Experience in Japan has shown that in the absence of 'rich diet' there is no CHD epidemic, even with high rates of smoking. The risk of coronary heart disease increases substantially when several factors are present. It is now acknowledged that coronary heart disease is a multifactorial process, with no one factor strictly determinative, essential or sufficient alone to produce the disease. In every instance, the risk associated with any factor has been found to vary according to the combination of other risk factors present. Thus a combined approach to the risk factors would seem to be sensible since they are frequently found clustered together. The major risk factors appear to be aetiological because they are strong and dose-related, predictive in a variety of population samples, independent of other risk factors, pathogenetically plausible and supported by clinical investigations (Inter-Society Commission for Heart Disease Resources, 1984). For example, the strength, consistency and graded nature of the relation between plasma

cholesterol and mortality from CHD make any explanation of the link other than a causal one extremely unlikely (Marmot, 1994). The size of the association between cholesterol and CHD mortality becomes even larger after correction for regression dilution bias, which arises from the random fluctuation of serum cholesterol concentration in subjects over time, and the surrogate dilution effect, which arises from the close relationship between total cholesterol concentration and low-density lipoprotein (LDL) cholesterol concentration (Law, Wald, Wu et al, 1994). The following sections will discuss each of the main CHD risk factors in greater detail.

1.1.2 Cholesterol

At all ages, the atherogenic potential of serum total cholesterol has been shown to derive from the LDL cholesterol fraction which is positively related to CHD incidence (Kannel, 1983). LDL cholesterol and high-density lipoprotein (HDL) cholesterol are the 2 main components of total serum cholesterol. LDL normally accounts for between 60 and 80 per cent of total serum cholesterol. It picks up cholesterol in the gut and dumps it in tissue, leading to a build-up of deposits. LDL cholesterol is known to transfer from plasma to the arterial wall at a rate which is directly related to its plasma concentration (Lewis, 1992). LDL cholesterol can be lowered slightly by diet, but large decreases require drug intervention. HDL normally accounts for between 15 and 25 per cent of total serum cholesterol, and it's role is to pick up cholesterol and transport it to the liver, so that it is flushed out of the body. HDL cholesterol has been shown to be inversely related to CHD incidence, which is consistent with it's metabolic role in removing cholesterol from the tissues (Kannel, 1983). HDL levels can be increased by exercise (Goldberg, 1989). Reflecting this 2-way traffic in cholesterol, the ratio of total to HDL cholesterol has been shown to be an efficient and convenient measure of lipid risk profile (Kannel, 1983) and a strong predictor of mortality (Goldbourt et al, 1985). The remaining component of total serum cholesterol is very-low-density lipoprotein (VLDL) cholesterol. Triglyceride is another important lipoprotein. It makes up most of the dietary fat consumed in industrial nations, and has been shown to be an independent risk factor for CHD in women, although it is not such a powerful predictor in men,

particularly after adjustment for HDL cholesterol levels with which it is correlated (Thelle, 1991).

Total serum cholesterol levels also rise with age. For women, the gradient of this rise increases sharply around age 35, while for men the slope levels off around age 45, as can be seen from Figure 1.2.

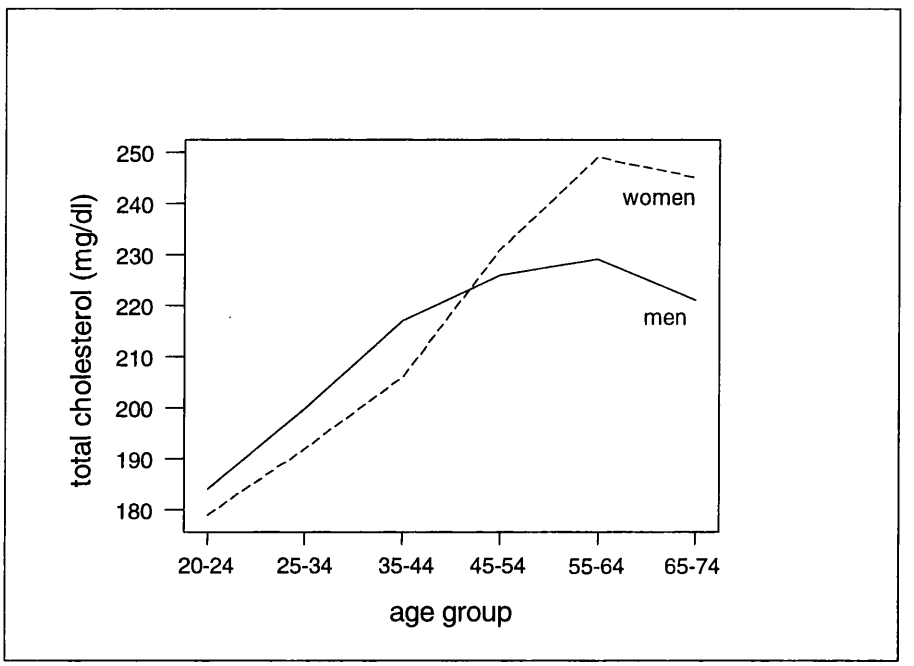


Figure 1.2 Mean total cholesterol levels according to age and gender (Data from the National Health and Nutrition Examination Survey II (Fulwood et al ,1986)).

In general, women have higher concentrations of HDL cholesterol than men (10 mg/dl or 0.26 mmol/l higher on average) which change little with age (Harlan and Manolio, 1992), and lower concentrations of LDL cholesterol until the menopause. This leads to women being exposed to cholesterol risk later in life than men, with coronary heart disease occurring an average of 7-10 years later in women than in men (Hulley et al, 1992) (see also the death rates for 45-64 year old men and women in Figure 1.1). After the menopause, the combination of rising LDL cholesterol and unchanging HDL cholesterol creates a less favourable risk profile for women. Post-menopausal oestrogen

replacement therapy has been shown (in the Lipid Research Clinics Program Follow-up Study) to increase HDL cholesterol and lower LDL cholesterol, leading to 40-50% fewer CHD events and lower overall mortality (Bush et al, 1987) but there may be other risks associated with this therapy.

Published studies vary in their units of cholesterol measurement. For reference:

$$1 \text{ mg/dl} = 0.026 \text{ mmol/l}$$

The British Hyperlipidaemia Association has issued guidelines on strategies for reducing coronary heart disease, and desirable limits for blood lipid concentrations (Shepherd et al, 1987). The recommended total cholesterol concentration is one below 5.2 mmol/l (200 mg/dl). Any subject in the general population with cholesterol level > 5.2 mmol/l should receive dietary advice, >6.5 mmol/l (250 mg/dl) should receive clinical supervision and possibly drug intervention, although drugs will be required mainly by patients with cholesterol level > 7.8 mmol/l (300 mg/dl), most of whom will suffer from familial hypercholesterolaemia.

Cholesterol measurement alone would not make an optimal screening test to detect those at high risk of CHD because of the substantial overlap between the relative frequency distributions of cholesterol levels for those who do and do not develop CHD (Pooling Project Research Group, 1978). Stress, dietary variation, acute illness, seasonal effects, posture and aspects of the blood sampling technique can all cause changes in cholesterol levels (Gordon et al, 1987; Hegsted and Nicolosi, 1987). It has been estimated that for cholesterol measurements taken one year apart, with no concurrent lipid intervention, the within-person coefficient of variation is 7%, compared with a between-person coefficient of variation of 15% (Thompson and Pocock, 1990).

1.1.3 Smoking

Cigarette smoking is also a major risk factor for coronary heart disease although it may be of less primary importance than dietary factors. The Seven Countries Study (Keys, 1980) found that smoking was a strong risk factor for CHD mortality and incidence in the USA, but only a weak risk factor in Japan. The strength of smoking as a risk factor

may be related to the background level of risk as determined by diet and levels of plasma lipids. Cigarette smoking has a greater influence in men than women and it's effects appear to be reversible on cessation (Kannel et al, 1984).

1.1.4 Fibrinogen

It is said that about half of the risk associated with cigarette smoking could be attributed to higher levels of plasma fibrinogen. Each standard deviation increase in fibrinogen has been found to be associated with a 1.6-fold increase in CHD incidence, a risk ratio close to that observed for cholesterol (Kannel, 1992). Fibrinogen and factor VII are two of the main determinants of the formation of clots, and thus influence thrombosis. They have been shown to be strong and independent risk factors for CHD in middle-aged populations (Harlan and Manolio, 1992; Kannel et al, 1987). High fibrinogen levels increase viscosity, enhance platelet aggregability and contribute to the development of atheroma, at least in its more advanced stages (Meade, 1987). The main environmental determinant of fibrinogen levels is smoking, with which there is a dose-response relationship (Wilkes et al, 1988), although fibrinogen levels are also associated with increased CHD risk in non-smokers (Meade et al, 1986). Thrombosis-related mechanisms can also be favourably influenced by dietary intake of fish oil. Fish oil also affects serum lipids, with the greatest changes occurring in the lowering of triglyceride levels (Elwood, Burr and Sweetman, 1992).

1.1.5 Diabetes

The relative risk of CHD incidence in diabetics compared to non-diabetics has been found to be higher in women than in men (Kannel, 1985), so that, for diabetics, the sex differences in CHD risk are virtually abolished. In men, this relative risk of CHD for diabetics compared to non-diabetics has been estimated as 2.4, while in women it was 5.1. Risk levels become even higher if the diabetes is insulin-dependent.

1.1.6 Obesity

Obesity, defined as an excess of body fat, frequently results in a significant impairment to health. The most simply measured and widely used index of body fat is Body Mass Index, BMI = $\frac{\text{weight}}{\text{height}^2}$. This is also known as Quetelet's index and is strongly correlated with more direct estimates of body fatness such as body density (Keys et al, 1972) or total body potassium (Larsson et al, 1981). Studies of large cohorts have found a U-shaped relationship between obesity and all-cause mortality (Larsson, 1992) with the optimum BMI being between 22 and 29 kg/m^2 for middle-aged subjects. Thinness is associated with mortality from obstructive lung disease, tuberculosis and stomach and lung cancer (Waaler, 1983), while the association between obesity and longevity has shown conflicting results due to short follow-up and small studies (Manson et al, 1987). Most larger studies have found general obesity to be an independent risk factor for CHD in both men and women (Rissanen et al 1989; Manson et al 1990), although obesity is also important through its close links with the development of other risk factors over time, including an association between obesity and smoking, and the obvious relationship between obesity and cholesterol levels via diet. Some recent studies have suggested that abdominal obesity (sometimes estimated by abdominal skinfold thickness) may be more closely related to CHD than general obesity (Hartz et al, 1990), with relative risks for abdominal obesity indicating almost as strong a risk as for the major CHD risk factors.

1.1.7 Lack of Exercise

Vigorous physical activity, especially aerobic exercise, has been associated with lower CHD incidence in a wide variety of populations (Morris et al 1980; Donahue et al 1988; Leon et al 1987). Adequate, habitual aerobic exercise in leisure time improves cardio-respiratory fitness and performance and thus confers protection against the occurrence of CHD. Its effect mainly relates to the acute phases of the disease, such as thrombosis,

although it is also of some benefit in counteracting standard risk factors (Morris, 1992), for example its role in raising HDL cholesterol levels. (Goldberg, 1989)

1.1.8 Blood Pressure

Despite the strong independent effects of blood pressure on CHD incidence, controlled trials have been inconclusive (MacMahon et al, 1986). This may be due to the antihypertensive agents used having adverse effects on other CHD risk factors, which cancel out the benefits of reduced blood pressure (Poulter, 1991). Exposure to blood pressure risk can also be influenced by salt intake, obesity and alcohol intake. It is thought that diastolic blood pressure is more closely related to cardiovascular disease before age 45, while systolic blood pressure is a better predictor after age 45 (Darne et al, 1989).

1.1.9 Alcohol

There has been some evidence from recent studies that alcohol consumption may have a 'protective' effect in lowering coronary heart disease risk. Hartung et al, (1983), suggested that this effect may be due to higher levels of high density lipoprotein in drinkers. This must however be balanced against the increased risk of death from other causes with high alcohol consumption. Alcohol raises HDL cholesterol, but this benefit may be offset by induced rises in blood pressure and triglyceride. There are problems with the measurement of alcohol consumption in that it may vary considerably from one point in time to another, which may have an effect on the predictive relevance of recorded consumption at any given point in time. The 'non-drinkers' group may also include people who have recently stopped drinking because of health problems, which will lead to bias and difficulty in interpreting associations.

1.1.10 Social Class

Social class has also shown interesting relationships with coronary heart disease. The Black Report (Black et al, 1988) concluded that social class differences in health are due to material conditions of life that are correlated with income levels, for example, nutrition and smoking. Further investigation is needed into risk factors which may account for socio-economic differences in health.

1.1.11 A look forward

This thesis will focus on the West of Scotland Coronary Prevention Study (WOSCOPS), (see Chapter 2). WOSCOPS has measured all of the above mentioned risk factors, although it's primary interest is in serum cholesterol. Cigarette smoking is an important risk factor which should always be considered, and alcohol has become of increasing interest in recent years. WOSCOPS also included a measurement of fibrinogen, currently emerging as a potentially important risk factor. Investigation into relationships among the above risk factors and various disease outcomes will be carried out primarily using the large cohort of subjects screened for WOSCOPS.

1.2 Evidence from epidemiology

1.2.1 The Framingham Heart Study

The Framingham Heart Study, one of the longest-running epidemiological studies, began in 1948 and has followed up a cohort of 5070 men and women, age 35-64 and living in the area of Framingham, Massachusetts, for over 30 years, with participants being examined every two years. It found (Stokes et al, 1987) that hypertension, total cholesterol, smoking and obesity are the major risk factors for cardiovascular disease.

Consideration of 3 sub-cohorts of the Framingham population, men age 50-59 at each of 1st January 1950, 1960 and 1970, showed (Sytkowski et al, 1990) that there has been a decline in the mortality from cardiovascular disease over the past 30 years. Comparison of the 1950 cohort with the 1970 cohort, for men who were free from cardiovascular disease at baseline, indicates that this decrease may be due to an improvement in the cardiovascular risk factors, with the 1970 cohort displaying lower serum cholesterol and systolic blood pressure, and reduced cigarette smoking. Analysis of 24 years of follow-up of the Framingham cohort (Friedman and Kimball, 1986) found a negative relationship between alcohol consumption and coronary heart disease for all males, and for female smokers, with no relationship for female non-smokers. In non-smokers, beer and wine consumption showed greater reductions in coronary heart disease mortality than consumption of spirits.

1.2.2 The Whitehall Study

The Whitehall Study examined mortality in 10 years of follow-up from initial screening between 1967 and 1969, for 17,530 office-based male civil servants in London. This study found a steep inverse relation between civil service employment grade and mortality. Men in the lowest employment grade had 3 times the mortality rate from CHD as men in the highest grade (Marmot et al, 1984). Smoking and other coronary risk factors are more common in the lower grades, but they account for only part of the difference in mortality. Consideration of the 1422 men in this study who had completed dietary records, revealed a U-shaped relationship between alcohol consumption and mortality (Marmot et al, 1981). This relationship was just as strong after excluding the first 2 years of follow-up, making it less likely that the high mortality among 'abstainers' was due to people giving up alcohol because they were already sick. Exclusion of the first two years of follow-up removed the U-shaped relationship between total mortality and plasma cholesterol, and left a positive relationship between them (Rose and Shipley, 1980).

1.2.3 The Multiple Risk Factor Intervention Trial

The Multiple Risk Factor Intervention Trial (MRFIT) required a large screening exercise, which involved over 300,000 men aged between 35 and 57 at baseline in 1973-1975, from 18 US cities (Kannel et al, 1986). Analysis of this large cohort confirmed the independent effects of serum cholesterol concentration, blood pressure and cigarette smoking as risk factors for coronary heart disease and all-cause mortality, although the strength of the associations diminished with increasing age. The relationships between age-adjusted CHD mortality and quintiles of serum cholesterol and systolic blood pressure (SBP) and baseline cigarette smoking are shown in Table 1.1. These indicate that CHD increases with serum cholesterol for both smokers and non-smokers, and CHD increases with SBP quintile at every level of cholesterol for both smokers and non-smokers. A similar pattern appears with diastolic pressure (Stamler et al, 1989).

	Systolic Pressure				
Serum Total Chol (mg/dl)	<118	118-124	125-131	132-141	>142
Non-smokers					
<182	3.09	3.72	5.13	5.35	13.66
182-202	4.39	5.79	8.35	7.66	15.80
203-220	5.20	6.08	8.56	10.72	17.75
221-244	6.34	9.37	8.66	12.21	22.69
>244	12.36	12.68	16.31	20.68	33.40
Smokers					
<182	10.37	10.69	13.21	13.99	27.04
182-202	10.03	11.76	19.05	20.67	33.69
203-220	14.90	16.09	21.07	28.87	42.91
221-244	19.83	22.69	23.61	31.98	55.50
>244	25.24	30.50	35.26	41.47	62.11

Table 1.1 Baseline cigarette smoking, quintiles of serum cholesterol, systolic pressure and age-adjusted CHD mortality per 10,000 person-years for men screened for the MRFIT.

A distinct escalation of risk was noted for combinations of these three risk factors. It was estimated that elimination of these risk factors had the potential for reducing the coronary heart disease mortality rate by two thirds in 35-45 year old men, and by half in 46-57 year old men. The relationship between coronary heart disease death and serum cholesterol quintile observed in the MRFIT study is illustrated in Figure 1.3 (Kannel et al, 1986), which also shows that risk increases with age.

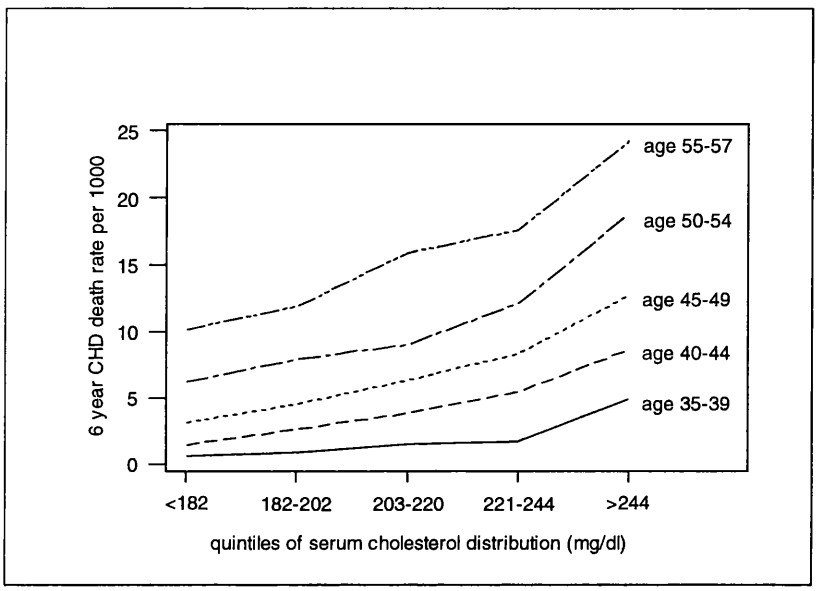


Figure 1.3 Relationship between coronary heart disease death and serum cholesterol quintile for various age bands observed in the MRFIT.

Further analysis of the MRFIT screening data (Martin et al, 1986) showed that above the 20th percentile for serum cholesterol (>181 mg/dl, or >4.68 mmol/l), coronary heart disease mortality increased progressively. Using men below the 20th percentile as a baseline risk group, half of all coronary heart disease deaths were associated with raised serum cholesterol concentrations, and half of these excess deaths occurred in men with cholesterol levels above the 85th percentile (>253 mg/dl, or >6.54 mmol/l). Diastolic blood pressure had a risk curve shaped similarly to cholesterol, for both coronary heart disease and total mortality (that is, a positive slope for the relationship with CHD and a J-shaped relationship with total mortality).

1.2.4 The British Regional Heart Study

HDL cholesterol and triglycerides were examined in relation to total serum cholesterol in the British Regional Heart Study (BRHS) (Pocock et al, 1989). This risk factor survey involved 7735 men aged 40-59, between 1978 and 1980, selected from general practices in 24 British towns chosen to reflect regional variations in CHD, including 3 in Scotland (Ayr, Dunfermline and Falkirk). 99% of these men were traced to their current GP between 9.5 and 11 years after screening, for follow-up purposes. Data collected included a standard questionnaire, physical measurements, blood samples, ECGs and measurements of respiratory function (Shaper and Elford, 1992). The BRHS found that, after adjusting for HDL and other risk factors, men in the highest quintile of the total cholesterol distribution were at 3.5 times the risk of ischaemic heart disease as men in the lowest quintile. Conversely, men in the lowest quintile of HDL concentration were at twice the risk of men in the highest quintile after adjusting for total cholesterol concentration and other risk factors. Triglycerides did not have a predictive importance once other risk factors had been taken into account. Consideration of place of birth information revealed that geographic zone of examination was a more important risk factor for CHD than zone of birth (Elford et al, 1989). Inter-town variation in CHD mortality was found to be associated with mean blood pressure, % of men with hypertension, current cigarette smoking, heavy drinking and the % of manual workers (Shaper and Elford, 1992). This study also found that ischaemic heart disease mortality rates were higher in manual than in non-manual workers (Pocock et al, 1987), with much of this increased risk being due to differences in cigarette smoking. Manual workers also had higher blood pressure, were more obese and took much less physical activity in leisure time. Although adjustment for these factors narrowed the gap between the two groups, manual workers still had a 24% excess of ischaemic heart disease events.

1.2.5 The Scottish Heart Health and Scottish MONICA Studies

Scotland has a reputation for very high coronary heart disease mortality rates. Data from the World Health Organisation (WHO), on coronary heart disease mortality from 1969 until 1981 in North America, Japan, Europe and Australasia, showed (Figure 1.4) that Scottish rates were among the highest, although the rate for males has been falling (Tunstall-Pedoe et al, 1986; and Figure 1.1). It also showed that, in the 40-69 year old age band, female rates correlate very strongly with male rates for the same country, although they are only 20-30% of the male rates (see Figure 1.1).

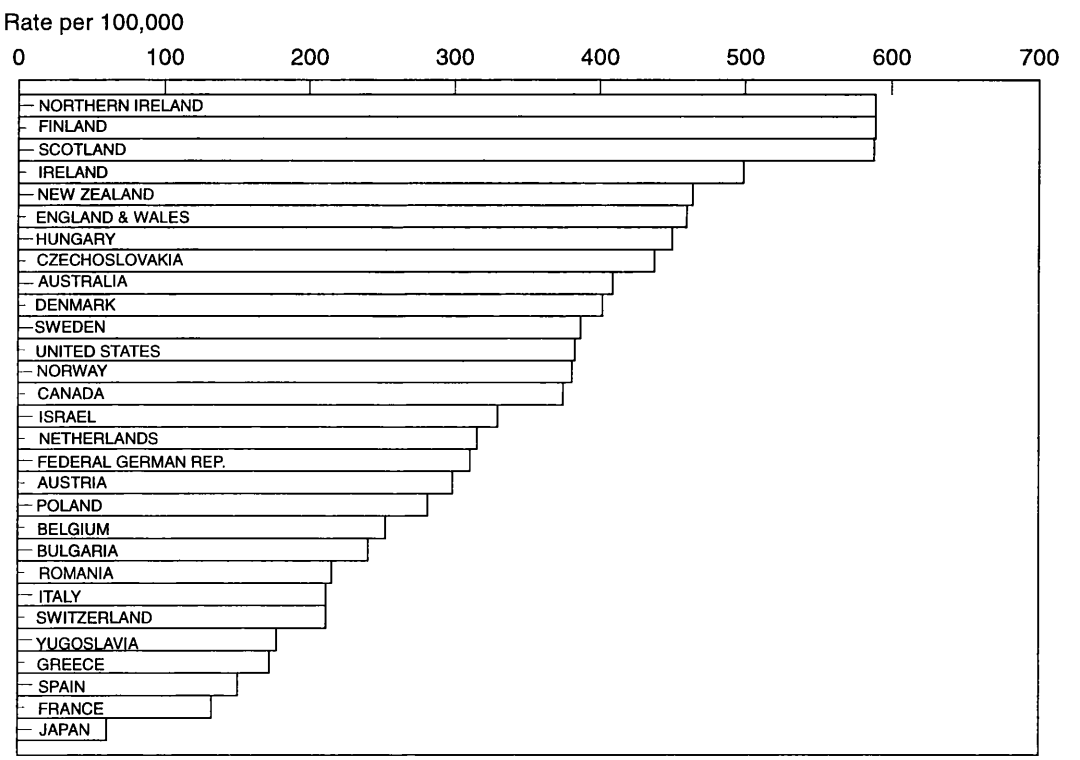


Figure 1.4 Age-standardised coronary heart disease mortality rates (per 100,000) for males age 40-69 (from Tunstall-Pedoe et al, 1986).

The Scottish Heart Health Study (SHHS) (Smith et al, 1989), a study of lifestyle and coronary heart disease risk factors measured between 1984 and 1986, in 10,359 men and women aged 40-59 in 22 of the 56 local government districts of Scotland, found that high cholesterol and cigarette smoking provided a classical explanation for the excess of coronary deaths in Scotland, but other factors such as dietary deficiencies merited further investigation. Considering one of the less frequently studied CHD risk factors, the SHHS observed that fibrinogen levels in men and women increased with age and were slightly higher in women than in men at each age group as shown in Table 1.2 (Smith et al, 1989). There was a weak association observed between fibrinogen levels and district smoking levels (Tunstall-Pedoe et al, 1989).

Age (years)	Mean values g/l (s.d.)	
	Men	Women
40-44	2.18 (0.62)	2.22 (0.64)
45-49	2.27 (0.66)	2.30 (0.62)
50-54	2.35 (0.73)	2.46 (0.69)
55-59	2.43 (0.71)	2.52 (0.74)

Table 1.2 Fibrinogen levels in the Scottish Heart Health Study

The main objectives of the SHHS were (Smith et al, 1987) to establish levels of coronary heart disease risk factors, explain geographical variation and assess the relative contribution of the risk factors, in the Scottish context, as well as to explain why Scotland as a whole has such a high coronary heart disease mortality rate. The SHHS found (Tunstall-Pedoe et al, 1989) that coronary mortality rates over 5 years varied between districts in Scotland by a factor of 2 in men and 3 in women, with lower rates generally being in the East, and higher rates in the West of Scotland. The SHHS observed an association between coronary heart disease mortality and % male unemployment, % in social class 3-5 and rainfall, explaining 73% of the geographical variation in mortality (Crombie et al, 1989). The SHHS followed the WHO protocol and

also included subjects in the age range of 40-59 from the WHO MONICA study, from Edinburgh and Glasgow, bringing the cohort size to approximately 12,000.

The aim of the MONICA Project was to measure the trends in and determinants of cardiovascular disease over 10 years in many different populations (WHO MONICA Project Principal Investigators, 1988). It's original design placed more emphasis on longitudinal measurement within each centre than on the comparability among centres. It involved 41 collaborating centres world-wide, resulting in a population of around 15 million men and women aged 25-64. In the Scottish branch of the MONICA Project, population surveys of coronary risk factors took place simultaneously in Edinburgh and north Glasgow in 1986. Comparison of these cohorts found that Glasgow in the west had a much higher cardiovascular mortality rate than Edinburgh in the east. This was related to the fact that major coronary risk factors were higher in north Glasgow than in Edinburgh, with the exception of serum lipids which were not significantly different (Smith et al, 1990). Again, with the exception of serum lipids, the risk factors were related to socio-economic status, so that most of the differences between the two cities disappeared after adjustment for housing tenure. So the socio-economic differences in coronary heart disease and its major risk factors may explain the apparent differences in CHD mortality rates between east and west Scotland.

1.2.6 Other studies looking at alcohol consumption

The findings in the Framingham cohort (see section 1.2.1) and the Whitehall study (see section 1.2.2) were in agreement with a review of ischaemic heart disease deaths in 18 developed countries (St Leger et al, 1979) which found a strong negative association between ischaemic heart disease deaths and alcohol consumption which was wholly attributable to consumption of wine. This effect may be due to trace components found in wine.

However, examination of the Malmo birth cohort of 4571 men born between 1926 and 1929 and screened between 1974 and 1978 (Petersen et al, 1980) found that heavy alcohol consumption was the single most important risk factor associated with

premature death from any cause in these middle-aged Swedish men, with serum cholesterol being an inverse risk factor.

In Britain, this relationship was examined via a postal questionnaire, sent to surviving male doctors (12,321 of whom replied) in 1978, with mortality follow-up until 1991. These men had originally been involved in the study into the effects of smoking which began in 1951. This study considered 7 categories of alcohol consumption and found a U-shaped relationship between alcohol consumption and all-cause mortality, with the lowest risk group being consumption of 8-14 units/week (Doll et al, 1994). However, all-cause mortality increases with amount drunk above 21 units per week.

The Albany Study (Gordon and Doyle, 1986) carried out in 1971-72 on 1910 civil service employees who had been enrolled into an earlier study in 1953-55 found that alcohol consumption was positively related to cigarette consumption, blood pressure and HDL cholesterol (although cigarette consumption was negatively related to HDL cholesterol). Thus the apparent alcohol effect may depend on other concurrent behaviour.

1.2.7 Overview Studies

Simons (1986) examined the coronary artery disease mortality rates for 19 countries, whose rates had previously been published by WHO. He found that, for men, 45% of the interpopulation variation in coronary artery disease mortality could be explained by interpopulation differences in total serum cholesterol levels, 32% by the variation in HDL cholesterol, and 55% by variation in the ratio total cholesterol/HDL cholesterol. For women, total cholesterol/HDL cholesterol was the only significant correlate of the coronary artery disease mortality rate, and explained 31% of the variation.

A more recent analysis of prospective cohort studies (Law, Wald and Thompson, 1994) showed that the size of the decrease in coronary heart disease incidence as cholesterol decreased was related to age. There have been at least 60 cohort studies of serum cholesterol and CHD, but this analysis has been restricted to the 10 largest published studies, which together involved 494,804 men. 7 of these studies considered only CHD

mortality, but 3 also considered non-fatal infarcts. A 0.6 mmol/l decrease in cholesterol (approximately 10% decrease) resulted in a decrease in coronary heart disease incidence of 54% on average, at age 40, but an average decrease of only 39% at age 50 and 27% at age 60. A cholesterol decrease of up to 0.6 mmol/l could be achieved by moderate dietary change without resorting to drug intervention which could reduce serum cholesterol by up to 1.2 mmol/l (approximately 20%).

1.2.8 Summary

These epidemiological studies support the relationships between CHD and the various risk factors described in section 1.1. They show that CHD risk is greater with raised cholesterol levels, high blood pressure and cigarette smoking, and have conflicting results with regard to the association between CHD and alcohol consumption. The Whitehall Study, the BRHS and the WHO MONICA Study have revealed interesting associations between socio-economic status and CHD which merit further investigation. The WOSCOPS study, which will be dealt with in this thesis will address these epidemiological issues through analysis of the cohort of subjects who came for initial screening.

1.3 Evidence from clinical trials

1.3.1 The Multiple Risk Factor Intervention Trial

The Multiple Risk Factor Intervention Trial (MRFIT) was a randomised primary prevention trial, involving approximately 12,000 35-57 year old men who were in the upper 10-15% of CHD risk. Screening for this study took place between 1973 and 1975, and it was designed to test the effect of a multifactor intervention program on mortality from coronary heart disease (MRFIT Research Group, 1982). The MRFIT was special in that it took a multiple factor approach, while the vast majority of clinical trials involve

only unifactor intervention. The 'special intervention' group received counselling for cigarette smoking, treatment for hypertension and dietary advice for lowering blood cholesterol. The control group relied on their usual sources of health care in the community. Follow-up was for an average of 7 years, but the difference in CHD mortality between the two groups was not statistically significant (there were 17.9 deaths per 1000 in the intervention group and 19.3 deaths per 1000 in the control group). It has been suggested that this may have been due to the decline of risk factor levels in both groups, resulting in a lower than expected mortality in the control group, or to unfavourable responses to antihypertensive drugs in some subjects in the treatment group.

1.3.2 Primary Prevention Trials of Cholesterol Lowering drugs

The three largest primary prevention trials of cholesterol reducing agents are described below.

The WHO Clofibrate Trial was started in 1965 to test the hypothesis that ischaemic heart disease incidence in middle-aged men could be reduced by lowering raised serum cholesterol levels. It involved 15,745 men from 3 European centres - Edinburgh, Budapest and Prague, approximately 10,000 of whom fell in the upper third of the cholesterol distribution determined from the 30,000 screened volunteers, and approximately 5000 from the lower third to be used as a second control group. The men with raised cholesterol were randomised to either clofibrate or placebo (Committee of Principal Investigators, 1978). A mean reduction of approximately 9% of the initial serum cholesterol values was achieved in the treatment group, resulting in a 20% decrease in the incidence of ischaemic heart disease, compared to the high cholesterol controls. The low cholesterol group showed substantially lower rates of ischaemic heart disease than either of the high cholesterol groups. The trial thus supported the hypothesis that lowering high serum cholesterol could reduce the incidence of ischaemic heart disease. However, this drug was not recommended for community-wide primary

prevention because of concerns regarding adverse reactions to the drug clofibrate. A further four years of follow-up until the end of 1982, giving a total mean follow-up of 13.2 years with a mean of 5.3 years in the trial and 7.9 years beyond the end of the trial revealed that the excess mortality in the clofibrate treated group did not continue after the end of treatment. There was an excess mortality of 47% during treatment, but only 5% after treatment ended. The authors were unable to explain the substantial excess of non-coronary mortality during the trial (Committee of Principal Investigators, 1984).

The Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) was a multi-centre randomised trial of the cholesterol lowering drug cholestyramine vs. placebo. Screening took place from 1973 to 1976 in 12 lipid research clinics across the US, to identify 3806 middle-aged men with primary hypercholesterolaemia, who also followed a moderate cholesterol lowering diet over the average of 7.4 years follow-up for the trial. The treatment group showed a 24% reduction in definite coronary heart disease death and a 19% reduction in non-fatal myocardial infarction in conjunction with an average 13.4% reduction in plasma total cholesterol and a 20.3% reduction in LDL cholesterol. These reductions were 8.5% (for total cholesterol) and 12.6% (for LDL cholesterol) greater than those obtained in the placebo group, which was also on the cholesterol lowering diet (Lipid Research Clinics Program, 1984, Part I). The cumulative seven year incidence of coronary heart disease death or non-fatal myocardial infarction was 7% in the cholestyramine group and 8.6% in the placebo group. The active treatment group also had lower rates of new positive exercise tests, angina and coronary bypass surgery. Small increases in HDL cholesterol which also accompanied cholestyramine treatment independently accounted for a 2% reduction in CHD risk (Lipid Research Clinics program, 1984, Part II). The risk of death from all causes was not however significantly reduced in the treatment group, which had a greater number of violent and accidental deaths which were not thought to be related to cholestyramine therapy.

The Helsinki Heart Study was a randomised 5 year trial of the safety and efficacy of the cholesterol lowering drug gemfibrozil in dyslipidemic middle-aged men. It involved 4081 Finnish men aged 40-55 with high levels of cholesterol. The study began in 1980 and was completed in 1987. A reduction of 34% was observed in the incidence of

coronary heart disease events in men on gemfibrozil treatment. Compared with placebo, gemfibrozil produced average decreases of 10% in total serum cholesterol, 11% in LDL cholesterol and 35% in triglycerides, and a mean increase of 11% in HDL cholesterol over the 5 years of follow-up, with the lipid changes being related to lipid levels prior to treatment and compliance with the medication regime. The changes in HDL and LDL cholesterol levels were both significantly associated with the reduction in CHD incidence rates in the gemfibrozil group after correction for other CHD risk factors including age, blood pressure, smoking and drinking habits, baseline lipid levels, exercise and relative weight. This study suggested that elevating HDL cholesterol and lowering LDL cholesterol are both effective strategies in the primary prevention of CHD (Manninen et al, 1988). This study also found (Manninen et al, 1992) that in the placebo group the best single predictor of cardiac events was the ratio of $\frac{\text{LDL cholesterol}}{\text{HDL cholesterol}}$. In combination with the serum triglyceride level, this ratio revealed a high risk group of subjects (representing approximately 10% of the trial population) with ratio > 5 and triglycerides > 2.3 mmol/l, which had a relative risk of cardiac events of 3.8 compared to all the other subjects. This high risk group received most benefit from treatment with gemfibrozil, with a 71% lower incidence of cardiac events in the subgroup at high risk and on gemfibrozil than in the corresponding high risk subgroup on placebo.

1.3.3 Secondary Prevention Trials

A review of published studies carried out by Silberberg and Henry (1991) indicated that cholesterol lowering to prevent CHD death is of far greater benefit in secondary prevention than in primary prevention. The authors calculated that the risk reduction achieved by drug therapy in secondary prevention studies of cholesterol lowering was 3.2%, while in primary prevention studies it was only 0.1%.

One of the most recent secondary prevention trials was the Scandinavian Simvastatin Survival Study (4S). This study involved 4444 men and women aged 35-70 with a history of angina pectoris or acute myocardial infarction and a serum cholesterol level

between 5.5 and 8.0 mmol/l. These subjects were randomised to treatment with either placebo or simvastatin, an HMG-CoA reductase inhibitor, between May 1988 and August 1989. Follow-up was for an average of 5.4 years, ending on 1 August 1994. Simvastatin therapy produced mean decreases of 25% in total cholesterol and 35% in LDL cholesterol, and a mean increase of 8% in HDL cholesterol. The relative risk of coronary death in the simvastatin group relative to the placebo group was 0.58, while for all-cause mortality the relative risk was 0.7. There were no significant differences between the two groups in terms of non-cardiovascular deaths. The study thus shows that treatment with simvastatin is safe, and improves survival in CHD patients post myocardial infarction (Scandinavian Simvastatin Survival Study Group, 1994).

1.3.4 Overviews

A meta-analysis by Yusuf et al (1988), of 22 randomised trials (9 primary and 13 secondary prevention) found a highly significant 23% reduction in risk of non-fatal myocardial infarction and CHD death with cholesterol lowering. The authors also reported that this reduction was directly related to both the degree and duration of the intervention to lower cholesterol levels.

Gordon et al (1989) examined, in American subjects, the inverse relation between HDL cholesterol and coronary heart disease as reported in the BRHS (Pocock et al, 1986). They looked at four prospective American studies. Two of these studies, the Framingham Heart Study and the Lipid Research Clinics Program Mortality Follow-up Study, were population based studies, while the other two, the Lipid Research Clinics Coronary Primary Prevention Trial and the Multiple Risk Factor Intervention Trial, were randomised trials of middle-aged men at high coronary heart disease risk (only the control groups were analysed here). On adjusting for age, blood pressure, smoking, body mass index and LDL cholesterol there was found to be a consistent inverse relation between HDL cholesterol levels and coronary heart disease incidence rates in each of these 4 studies with, for example, a 1 mg/dl increment in HDL cholesterol being associated with a decrease in cardiovascular mortality of 3.7% in men and 4.7% in

women in the Lipid Research Clinics Program Mortality Follow-up Study. HDL cholesterol levels were found to be unrelated to non-cardiovascular mortality.

In an overview analysis of 16 randomised clinical intervention trials (a mixture of primary and secondary prevention, diet and drug trials) of cholesterol reduction (Holme, 1990) it was found that for every 1% decrease in cholesterol level, there was an estimated 2.5% decrease in coronary heart disease incidence. It was also found, predictably, that drug trials tended to more efficacy at cholesterol lowering than dietary trials, that efficacy was higher in secondary than in primary prevention trials, and that efficacy depended on the baseline cholesterol level. This review also showed that while a 1% decrease in cholesterol was effective in lowering CHD incidence, the reduction must be at least 8-9% to be effective in lowering total mortality.

1.3.5 Summary

Evidence from controlled clinical trials shows convincingly that reducing serum cholesterol levels by diet or drug treatment reduces the incidence of coronary heart disease (Katan, 1990). Some of these trials have found that while coronary heart disease death rates were lower, there was no significant difference in all-cause mortality between the treatment and control groups. In the LRC-CPPT this was thought to be due to the larger number of violent and accidental deaths in the treatment group. However, the actual numbers of violent deaths in the LRC-CPPT was small (11 in the treatment group and 4 in the placebo group) so the study had very low power to address this outcome. The 4S study is of particular relevance to the WOSCOPS study, which will be dealt with later, as it involves a drug belonging to the same family - the HMG CoA reductase inhibitors. This study indicated that there may be substantial benefit to coronary mortality from this cholesterol lowering therapy, and some benefit for all-cause mortality.

1.4 Concerns about Reduction in Cholesterol Levels

While the relationship between high cholesterol levels and coronary heart disease is well established, there has been much controversy about a possible relationship between low cholesterol levels and cancer. Some epidemiological studies and meta-analyses of clinical trials have detected this relationship, but others have found no significant relationship, leading to results which, overall, are conflicting and unclear. Although previous clinical trials have indicated a decrease in cardiovascular disease with cholesterol lowering, meta-analysis has indicated no decrease in total deaths in the treatment groups, with an increase in deaths due to suicide and other violence as well as to cancers. Evidence relating to these non-cardiac outcomes will now be considered.

1.4.1 Cancer and Cholesterol

1.4.1.1 Epidemiological Evidence

In 1981, Feinleib observed a fairly consistent inverse relationship between baseline cholesterol and subsequent cancer mortality in the Framingham Study. For all-cause mortality there was a U-shaped relationship (as illustrated in Figure 1.5), indicating that there was an optimal range of cholesterol values for which total morbidity and mortality might be minimised. He observed these relationships only for men, with no significant trends appearing for women.

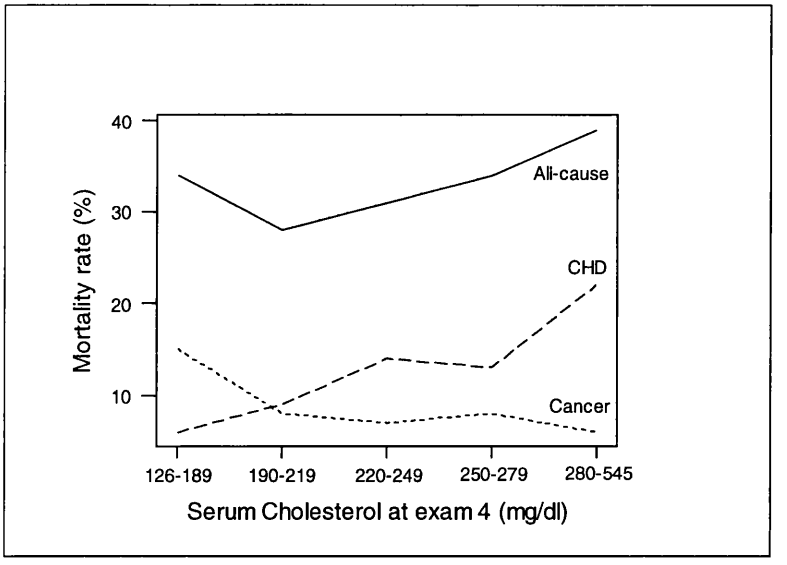


Figure 1.5 Age-adjusted mortality rates by serum cholesterol measured at examination 4 of the Framingham Study for men age 35-64.

A cohort of 5209 men and women from the Framingham study had repeat measurements of serum cholesterol made biennially, up to 18 years prior to cancer diagnosis (Sorlie and Feinleib, 1982). There was a significant inverse relationship, with the principal elevation in risk being for men with serum cholesterol less than 190 mg/dl. The hypothesis of pre-clinical cancer is refuted by the fact that an association of low cholesterol with incident cases occurred up to 18 years after baseline - most oncologists would be reluctant to accept that a metabolically active lesion would remain pre-clinical for 10 or more years, and the lack of a relationship between cholesterol and cancer in women - since it is reasonable to assume that once the cancer process has begun, the metabolic consequences for men and women would be similar. An analysis of time trends did not yield consistent results for all age groups. In some age groups, cholesterol levels were depressed 18 years before diagnosis, while in other age groups, the cholesterol level reduced further as the time of diagnosis approached. So low cholesterol may be both a precursor and a response to the cancer process.

The Lipid Research Clinics Program Mortality Follow-up Study involving 2753 men and 2476 women aged 40-79 at baseline between 1972 and 1976 and followed up until 1984 found that total cholesterol and LDL cholesterol were significantly inversely

associated with overall cancer mortality in men but observed no relation for women (Cowan et al, 1990). The relatively small numbers of cancer deaths restricted the site-specific analysis possible, but for colon cancer the relative risk of men in the lowest quartile of cholesterol levels (≤ 187 mg/dl) compared to the other three quartiles of cholesterol was 5.20, after correcting for age, body mass index, cigarette smoking and alcohol consumption. From consideration of the survival curves, the authors felt that the inverse relation observed in men was not due to pre-existing disease.

Salmond et al (1985) carried out a prospective epidemiological study on New Zealand Maoris with 17 years follow-up from first examination in 1962-63, and found a significant inverse relationship between serum cholesterol and cancer mortality for both men and women, after adjusting for age, systolic blood pressure and body mass index. This relationship remained significant after excluding deaths occurring in the first five years, indicating that the apparent relationship could not be explained by undetected illness.

Tornberg et al (1989) followed up 9217 Swedish men identified by screening between 1963 and 1965 for 18-20 years and also found a significant relationship between cholesterol and cancer incidence and mortality. However, in contrast to Salmond et al, they found that this relationship was strongest in the first two years of follow-up, consistent with the effect being due, at least in part, to preclinical cancer.

In the Honolulu Heart Program it was possible to consider several cancer sites separately. The Honolulu Heart Program examined the relationship between baseline serum cholesterol level and subsequent 9 year mortality in a cohort of around 8000 Japanese-American men age 46-65 who lived on the island of Oahu in 1965 (Kagan et al, 1981). There was a strong inverse relationship between serum cholesterol level and the subsequent risk of dying from cancer. This association persisted after excluding patients with evidence of disease at baseline, excluding deaths occurring in the first 6 years of follow-up, and adjusting for age, systolic blood pressure, smoking, alcohol consumption and relative weight. The only specific site which remained significant after these adjustments was cancer of the colon. The risk of death from cancer was nearly

four times greater in men with serum cholesterol less than 180 mg/dl than in men with serum cholesterol greater than 269 mg/dl. In contradiction, a further study on these Japanese-American men using 22 years of follow-up (Chyou et al, 1992) found that a significant negative association between serum cholesterol and smoking related cancers (lung, mouth, larynx, oesophagus, pancreas, bladder) did not persist after adjustment for cigarette smoking and alcohol consumption. The Honolulu Heart Program also found evidence of a U-shaped relationship between alcohol consumption and all-cause mortality (Blackwelder et al, 1980). While coronary heart disease decreased with increasing alcohol consumption, other disease outcomes such as cancer and stroke increased giving a total mortality picture as in Figure 1.6.

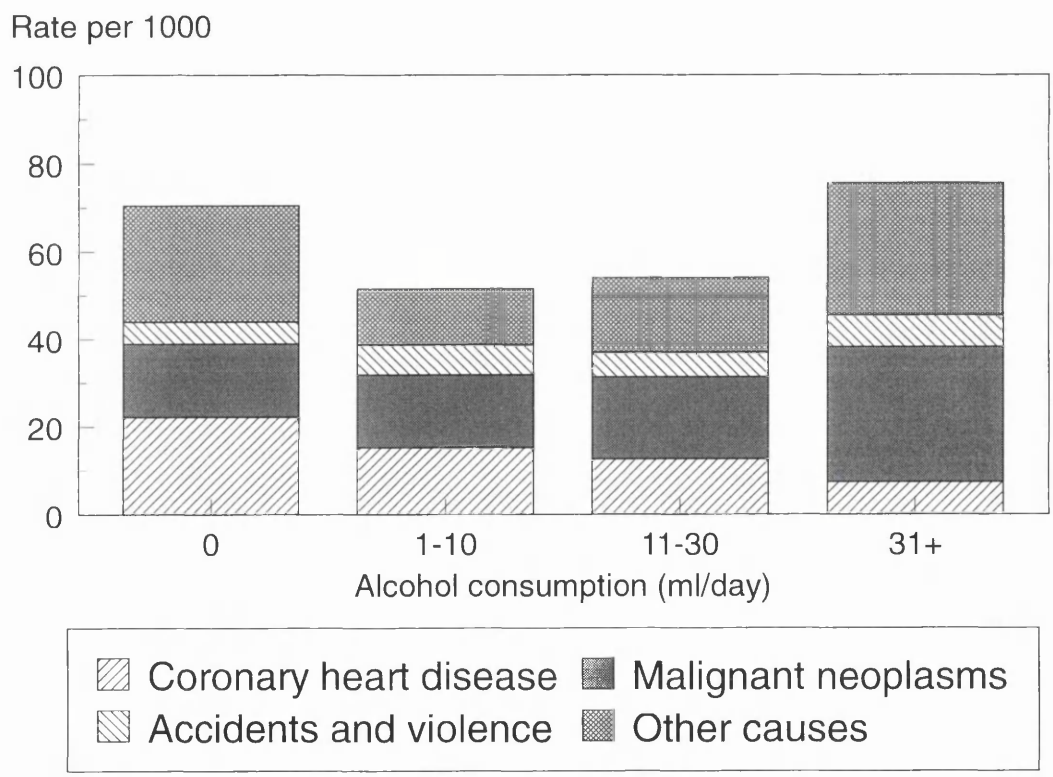


Figure 1.6 Age-adjusted 8-year mortality by cause and level of usual alcohol intake (adapted from Blackwelder et al, 1980)

A retrospective case-control study involving cancer patients, heart patients and controls, stratified for gender, age and smoking habit, and followed up for fourteen and a half

years for the Stockholm Prospective Study, with screening between 1970 and 1973 and follow-up back to 1963 and forward to 1982 (Gerhardsson et al, 1986), found an association between low serum cholesterol and cancer of the large intestine and rectum. However, this association was only apparent for a six-year period prior to death. Since this paper considers mortality rather than incidence, the low serum cholesterol levels are more likely to reflect an effect of preclinical cancer.

For the cohort of 300,000 men screened for the MRFIT between 1973 and 1975, mortality follow-up revealed a significant excess of cancer in the lowest decile of serum cholesterol level during the early years of follow-up, which decreased over time but had not disappeared after an average 7 years of follow-up (Sherwin et al, 1987). This finding is consistent with the inference that the association between low serum cholesterol and cancer is at least in part due to an effect of preclinical cancer.

During the same time period, a case-control study was carried out in Iowa with screening between 1973 and 1975 and follow-up until 1980 (Wallace et al, 1982). Controls were matched to cancer cases on age, sex, screening centre and date of screening, and the authors found that plasma cholesterol levels were lower among male cases and higher among female cases than among controls. The higher cholesterol levels in women were mainly for hormone-related cancers, while the lower levels in men were mainly for smoking-related cancers.

Meanwhile, in Scotland, the Renfrew and Paisley Survey (Isles et al, 1989), with screening from 1972 to 1976 and an average of 12 years of follow-up, found that the inverse association between cholesterol and cancer in men was strongest for lung cancer, unrelated to age, present for incidence as well as mortality, and persisted even when the first 4 years of follow-up were excluded. This study found no consistent relationship between cholesterol and smoking, counteracting another possible explanation. Out of 21 previous reports on cholesterol and cancer, examined by Isles et al, 8 showed no relationship, 12 found an inverse association and 1 found a positive association, which was with colorectal cancer.

A slightly different angle on the question was to consider a population with **naturally** low cholesterol levels. The mean cholesterol level in China is much lower than that usually observed in Western populations. A study of 9021 Chinese men and women age 35-64 in 2 screening cohorts (1972-73 and 1977-78) (Chen et al, 1991) revealed no significant association between baseline cholesterol and total mortality from cancer. There was however an inverse trend for liver cancer, excluding patients who died in the first three years of follow-up. There was also a significant inverse relationship between cholesterol and other chronic liver diseases for this population. The authors suggest that the lowered cholesterol levels were probably due to the fact that many Chinese suffer from long-term chronic hepatitis-B.

In the United States, a case-control study was carried out at the Mount Sinai hospital in New York City between 1978 and 1980 to examine cholesterol levels in relation to colon cancer (one of the more common cancers in the US). Controls were subjects who had entered the hospital for elective surgery. Comparison of 133 case-control pairs, matched by age and sex, revealed that colon cancer patients had significantly lower serum cholesterol levels than controls (Miller et al, 1981). However, examination of 130 case-case pairs in which early tumours were matched with advanced tumours found that women, although not men, had significantly lower cholesterol levels with advancing disease. This supports the idea that, in women at least, low serum cholesterol levels in cancer patients may be the result of metabolic influences of tumours.

The National Health and Nutrition Survey Epidemiologic Follow-up Study (Kritchevsky, 1992), which was conducted during 1981-84 using subjects who had previously been in the probability sample of US civilians used in the National Health and Nutrition Examination Survey from 1971-75, found that in females, the relationship between low cholesterol and cancer varied according to the amount of fat on the body. This suggests that, in women, a hormonal/metabolic basis may underlie the low cholesterol-cancer association. Diet could not, however, explain the low cholesterol-cancer association in men. In this study, the exclusion of early cancers strengthened the low cholesterol-cancer association, for males. This study also revealed (Schatzkin et al, 1988) that the inverse relationship was especially prominent for smoking related cancers, for which the association persisted 6 years after baseline

measurements, suggesting that the relationship cannot simply be dismissed as a pre-clinical cancer effect. There was no appreciable difference when analysis was carried out within strata of various risk factors such as age, education, poverty index, body mass index, smoking, alcohol consumption, race, dietary fat, fibre intake, and also, for women, age at first birth, age at menarche, parity and menopausal status.

1.4.1.2 Evidence from clinical trials

In the LRC-CPPT, it was observed (Kritchevsky et al, 1991) that the cholesterol levels of men diagnosed with non-localised cancer began to drop about 2 years prior to diagnosis. This decrease was not observed for men with localised malignancies. The decrease was 9.3 mg/dl on average, and weight also decreased by an average of 1.2 kg. Within eight months of diagnosis, both weight and cholesterol were significantly lower than expected. Kritchevsky et al recommended that analysis should exclude incidence within two years and deaths within three and a half years of baseline.

In the clinical trial branch of the MRFIT, it was found that the numbers of cancer deaths in the intervention and control groups were similar (81 in the intervention group and 69 in the control group) (MRFIT Research Group, 1982), suggesting that cholesterol lowering may not lead to excess cancer risk in subjects who start with high cholesterol levels.

Similarly to these primary prevention studies, the secondary prevention 4S study found no significant difference between the drug and placebo groups in terms of cancer deaths. There were 33 cancer deaths in the simvastatin treatment group, and 35 in the placebo group. In addition, there were 57 non-fatal cases of cancer in the simvastatin group and 61 in the placebo group (Scandinavian Simvastatin Survival Study Group, 1994).

1.4.1.3 Evidence from Overviews

In 1982, an International Collaborative Group examined 11 population studies from 8 countries, including the Renfrew and Paisley Survey in Scotland, and found a strong

inverse association between cancer and cholesterol in the first year of follow-up which diminished and disappeared in later years, consistent with the hypothesis that lower cholesterol levels were due to an effect of undetected disease.

This is consistent with an examination of 20 published cardiovascular disease studies, (McMichael et al, 1984), which afforded substantial evidence that preclinical cancer causes a lowering of blood cholesterol, and limited evidence that males with naturally low blood cholesterol levels are at increased risk of colon cancer. Deliberate lowering of blood cholesterol did not appear to alter the risk of cancer. It is possible that an individual's innate lipoprotein cholesterol profile is statistically associated with his/her predisposition to develop cancer, and the cholesterol-cancer relationship is thus not causal.

Law and Thompson (1991) analysed 33 prospective studies and discovered lower cholesterol levels in subjects diagnosed with cancer within 2 years of the baseline measurements. After this interval, the average differences in cholesterol levels were smaller, but still statistically significant, and could not be attributed to preclinical cancer. They concluded that the long-term association between low cholesterol and lung cancer is probably due to smoking, which may also explain how this association varied between studies according to the predominant socio-economic status of the subjects recruited. Law and Thompson found that the long term association was absent in studies of professional men, moderate in studies of mixed populations and strong in studies of manual workers. The association was confined to the first few years of follow-up in some of these 33 studies, but persisted despite many years of follow-up in others. There was an apparent association between high cholesterol and primary brain tumours. Cancer does act to lower serum cholesterol, to a certain extent, as does oestrogen, and the inverse association between cholesterol and cancer is much smaller than the direct association between cholesterol and coronary heart disease. So, overall, the authors argue that the data do not suggest that serum cholesterol reduction causes cancer.

The conference on low blood cholesterol (Jacobs et al, 1992), considered 19 cohort studies in a pooled epidemiologic study. The formal statistical overview adjusted for age, diastolic blood pressure, cigarette smoking, body mass index and alcohol intake as

available. Once again, they found a U-shaped relationship between cholesterol and all-cause mortality for men, reflecting the positive association with coronary heart disease and the negative relationship with cancer. There was no clear relationship for women due to the lack of power caused by the few numbers of women in the studies, and the results were unaltered by the exclusion of deaths occurring in the first 5 years of follow-up. It is suggested that the relationship between cholesterol and cancer may be due to some other confounding factor, although several factors were adjusted for in the analysis. It appears that the cholesterol-cancer relationship may be strongest in the lowest socio-economic group, but the authors felt that more research is needed in this area.

1.4.1.4 Discussion

From the above papers, it can be seen that there is much conflict in the findings of different studies regarding low cholesterol and cancer. Although some of the findings are from large studies, most come from smaller studies which have relatively little power on their own, or from meta-analyses. Also, most of the studies concentrate on middle-aged men, leading to an even less clear picture for women.

To summarise, some studies have found no relationship between cholesterol and cancer (for example, Chen et al 1991), some studies have found an inverse relationship which disappeared after the removal of deaths in the first few years of follow-up suggesting an effect of pre-clinical cancer (for example, Gerhardsson et al 1986), some studies found an inverse relationship which persisted after the removal of early deaths (for example, Sorlie and Feinleib 1982) [although there is considerable inter-study variation on how many years should be excluded from analysis], and some studies found a direct relationship, although this was mainly for hormone-related cancers in women (Wallace et al 1982).

It is plausible that the apparent relationship between cancer and low cholesterol may be due to preclinical cancer causing cholesterol lowering. A biological explanation for cholesterol lowering with pre-clinical cancer has been given in terms of an activation of

the membrane low density lipoprotein receptors of the cancer cells, so that cholesterol is incorporated into these cells, with a consequent reduction of its serum level (Budd and Ginsberg, 1986).

Buchwald hypothesised in 1992 that since cancer cells seem to demand an increase in cholesterol concentrations in themselves, cholesterol inhibition, either by decreased cholesterol availability or by decreased intracellular cholesterol synthesis, could inhibit tumour cell growth, act as an adjuvant to cancer chemotherapy and possibly prevent carcinogenesis. There are two sources of cellular cholesterol - synthesis within the cell and incorporation of extracellular cholesterol via receptor-mediated uptake of plasma LDL cholesterol. The cholesterol requirements of tumour cells exceed their total endogenous synthesis, so the tumour cells must use more cholesterol from the cell environment for their survival and growth. Buchwald (1992) found that, in rats, tumour weight was positively correlated with plasma cholesterol concentrations, and that, in mice, restriction of cholesterol synthesis using the HMG CoA reductase inhibitor lovastatin resulted in suppressed tumour growth. Buchwald's tissue-culture experiments point to an additive effect of cholesterol inhibition and chemotherapy on tumour cell growth, although this has yet to be tested in humans.

An alternative explanation of this relationship could be competing risks. Cancers comprise the most common contribution to non-cardiovascular mortality, and would thus be expected to show higher incidence at lower cholesterol levels where the cardiovascular risk is less (Williams et al, 1981). This argument of competing risks could be valid for elderly people who are followed up till death, but not for middle-aged subjects who are followed up for only a few years and most of whom are still alive.

This apparent relationship may also be due to confounding with other risk factors (such as smoking as a major risk factor for lung cancer), while some studies have suggested that the association between low cholesterol concentrations and cancer may be secondary to a relationship between low retinol (vitamin A) concentrations and cancer (Marenah et al, 1983).

However, there is a fear that it may be the act of cholesterol lowering which leads to increased risk of cancer, while indigenously low cholesterol has no adverse effects. The

possible relationship between low cholesterol and cancer is apparent in populations with high mean cholesterol levels, but there is little convincing evidence of high cancer rates in populations with naturally low cholesterol concentrations in countries where the diet is low in saturated fat and cholesterol, for example Japan (Katan, 1990). A further piece of evidence against cancer risk with indigenously low cholesterol is the lack of excess cancer mortality among people suffering from genetic conditions which cause LDL cholesterol levels of zero (although these conditions are relatively rare) (Katan, 1990). In the case of clinical trials, it is also possible that the cholesterol lowering drug used may be carcinogenic, although this is unlikely in general since the inverse trend between cholesterol and cancer has appeared with several different drugs (Oliver, 1992). There is, however, suspicion that the increase in cancer deaths was due to the particular drug involved in the large WHO clofibrate trial (Committee of Principal Investigators, 1984), which may, in turn, bias some of the meta-analyses carried out.

1.4.2 Suicide and other deaths

As mentioned earlier, the LRC-CPPT found a greater number of violent and accidental deaths in the treatment group compared to the controls, leading to no significant reduction in all-cause mortality for the treatment group, although there was a 24% decrease in definite coronary heart disease deaths (Lipid Research Clinics Program, 1984). However, the actual numbers involved were fairly small and the study was designed to look at coronary deaths, so it had insufficient power to address the endpoint of all-cause mortality.

The LRC-CPPT was considered with the Helsinki Heart Study (Wysowski and Gross, 1990), which also had higher rates of homicides, suicides and accidents in the treatment group than the placebo group. A detailed review of the individual case histories for each of these deaths provided little evidence to support the hypothesis that the cholesterol lowering drugs, gemfibrozil and cholestyramine, could be causally related to these deaths. The higher rates were not statistically significant in either study due to the small number of deaths involved.

In secondary prevention, the 4S study revealed no suggestion of a higher rate of violent deaths in the simvastatin treatment group. There were 6 violent deaths in the simvastatin group compared with 7 in the placebo group (Scandinavian Simvastatin Survival Study Group, 1994). Once more, there were very few events in this category of deaths.

The association between low cholesterol and violent deaths has appeared consistently in several primary prevention studies, whether cholesterol lowering was achieved by drugs or diet (Ryman, 1994). But this finding is not restricted to intervention trials. A significant association between low cholesterol and death from non-medical causes was also found in a population with naturally low cholesterol concentrations (Chen et al, 1991), suggesting that the association is due to the cholesterol level itself, and is not just a treatment effect (Ryman, 1994).

However, in contrast, a study carried out in Finland on two independent cohorts (one with baseline measurements carried out in 1972, and the other in 1977) with 10-15 years mortality follow-up, found no significant relationship between cholesterol levels and deaths from accidents, suicides and other violence (Vartiainen et al, 1994). This result may have been influenced by the relatively high cut-off point used for a 'low' cholesterol value in this study (< 5.0 mmol/l) (Ryman, 1994).
Examination of 25 years of follow-up for the two Finnish cohorts of the Seven Countries Study found a statistically non-significant negative association between cholesterol and mortality due to accidents and violence for one cohort, and a statistically significant positive association for the other cohort (Pekkanen et al, 1989). It is suggested that the observed associations between serum cholesterol and violent deaths are probably due to other, presently unknown, factors or to chance.

A meta-analysis of 6 primary prevention trials, including LRC-CPPT and HHS as above, found that while cholesterol lowering intervention reduced CHD mortality, it had no effect on total mortality (Muldoon et al, 1990). No consistent relation was found between reduction in cholesterol and mortality from cancer, but there was a significant increase in deaths not related to illness. In contrast to the group on cholesterol lowering therapy, mortality from suicides, accidents and other violent deaths in the control group was similar to the American national average. The near doubling of mortality from

violent causes observed in the group on cholesterol lowering therapy compared to the control group is inconsistent with the hypothesis of competing risks. The authors state that there is some evidence that modifying fat in the diet has both neurochemical and behavioural consequences, which may cause these excess deaths.

Considering these consequences of low cholesterol, a study of male mental hospital patients (Sletten et al, 1964) observed that low cholesterol subjects were more regressed and withdrawn, with less evidence of initiative and positive mental health than subjects with higher cholesterol levels. However, this gives no indication of whether low cholesterol is a precursor or a result. The California Psychological Inventory (Jenkins et al, 1969) found that impulsive behaviour traits were associated with low cholesterol levels. A study of 280 male homicidal offenders, carried out in Finland from 1974-1981 (Virkkunen, 1983), found that there was a clear relationship between low cholesterol and a habitually violent tendency under the influence of alcohol, and suggested that this may be due to an enhanced insulin secretion underlying both aspects. A smaller study, involving 47 adolescent boys (Virkkunen and Penttinen, 1984) found significantly lower cholesterol in boys with aggressive conduct disorder than in controls. It was again proposed that this may be due to the effects of active insulin in regulating LDL cholesterol.

An alternative biological explanation may come from an effect of lowering cholesterol on cerebral cell metabolism, leading to bizarre behaviour and poor co-ordination. In the central nervous system, serotonin suppresses harmful behavioural impulses (Engelberg, 1992). Low membrane cholesterol decreases the number of serotonin receptors, which may contribute to a decrease in brain serotonin and poorer suppression of aggressive behaviour. Abnormalities in cerebral serotonin systems are associated with poor impulse control, which may manifest itself either as suicidal behaviour or as aggression towards others. Long-term follow-up of middle-aged men may not show this relationship so clearly since those with a naturally low cholesterol may have already died in adolescence. A subsequent lowering would then be needed to trigger violence or suicide in older men, but this triggering may not be related to the magnitude of the reduction. It is not unreasonable that anything which affects the balance of central lipid metabolism could have profound effects on brain function.

Possible confounding variables which may influence the proposed relationship between cholesterol and violent death include alcohol consumption, blood pressure and social class. For example, in clinical trials, people may react differently to their usual alcohol intake, while taking drugs. Analysis of the Italian cohorts of the Seven Countries Study (Menotti et al, 1980 and 1987) revealed a direct relationship between blood pressure and violent death. Indeed, a review of specific cases found that some fatal accidents were preceded by minor cardio-circulatory events. These possible confounding variables may be worthy of further investigation.

A review of 10 cohort studies, 2 international studies and 28 randomised trials (Law, Thompson and Wald, 1994) found an excess in mortality with low or lowered serum cholesterol. The main cause of death attributable to low cholesterol was haemorrhagic stroke, with the excess risk being for cholesterol concentration less than 5 mmol/l. However, stroke affects only about 6% of Western populations, so this excess risk will be outweighed by the benefits from the low risk of ischaemic heart disease. Employed cohorts showed no excess non-circulatory mortality, while community cohorts showed cholesterol associations with suicide, chronic bronchitis and chronic liver and bowel diseases, as well as lung and haemopoietic cancers. These excess deaths can mostly be explained by early disease, or by other factors, such as depression leading to suicide, and also causing cholesterol lowering. This suggests that the lowering of cholesterol concentrations causes an increase of deaths only in a population more vulnerable to factors such as psychiatric disorder, and that employed cohorts are protected from this effect (Ryman, 1994). It is argued that for a relationship to be causal, it must also be reversible. Treating depression leads to an increase in serum cholesterol, suggesting that it is the depression which causes the low cholesterol. (Similarly, cholesterol increases when chemotherapy induces remission of cancer, backing the pre-clinical cancer argument.) Low cholesterol seems to be associated with suicide only for the first 5 years after measurement. Cholesterol can predict ischaemic heart disease deaths 30-40 years later and should do the same if it were a cause of suicide. Alcoholism is also associated with depression and other mental illness, causes deaths through accidents and suicides, and lowers serum cholesterol. So there appears to be some evidence for the safety of lowering serum cholesterol, as the adverse effects may be explained by confounding factors, or as having a cholesterol lowering effect themselves.

Excess stroke deaths with low cholesterol concentrations were found in an analysis of the MRFIT screening exercise, which involved 6 years of follow-up on 350,977 men age 35-57 (Iso et al, 1989). Deaths from intracranial haemorrhage were three times higher in men with serum cholesterol levels under 4.14 mmol/l than in those with higher levels, but there was also a positive association between serum cholesterol and deaths from non-haemorrhagic stroke. The deaths from intracranial haemorrhage were confined to men with diastolic blood pressure ≥ 90 mmHg, among whom death from intracranial haemorrhage is relatively common. The risk of haemorrhagic stroke death with low cholesterol is thus overwhelmed by the risks of non-haemorrhagic stroke and coronary heart disease with high cholesterol levels.

Analysis of 17 years of follow-up in the Honolulu Heart Program also found significant inverse trends between cholesterol and deaths due to haemorrhagic stroke, all cancer, benign liver disease and chronic obstructive lung disease (Frank et al, 1992). Only the inverse trends for all cancer and benign liver disease flattened when deaths in the first five years of follow-up were excluded.

1.5 Summary and Proposed Investigation

As has been seen above, while cholesterol is well established as a risk factor for coronary heart disease, it's effects on cancer, suicide and other violence are much less clear. The studies have conflicting results, making this a controversial area of medical research.

Differences in study design - whether it is prospective or retrospective, the cancer outcome used (incidence or mortality), the sample size involved in the study, the average cholesterol levels in the different population groups used, the age of patients at the end of the study and the ethnic composition of the study subjects could all contribute to some of the inconsistent findings between studies. The selection of the population entering the study has the potential to be a major source of bias. It is also important to note the length of follow-up in the different studies, especially when considering the

possible relationship between cancer and cholesterol, where preclinical disease could bias results in the first few years of follow-up. A further problem is that the studies described above were designed to examine cardiac endpoints and do not have sufficient power to address alternative outcomes. This is particularly relevant for the outcome of 'violent deaths' where the actual numbers of deaths observed is small. Differences in results may also be due to differences in the methods of analysis employed (Salmond et al, 1985). The studies also vary with respect to which covariables are adjusted for in the analysis. In meta-analyses, drug and dietary trials with varying periods of follow-up and degrees of success are often considered together. Amidst the conflicting findings it is difficult to discern whether the observed trends relate to cholesterol lowering in general, to specific drugs for cholesterol lowering, to naturally low cholesterol or are simply a statistical artefact.

Lowering of plasma cholesterol should be achieved by diet alone, where possible, with drug intervention being reserved only for those at very high risk of coronary heart disease. Cholesterol reduction should be desirable only for those with raised cholesterol concentrations, and not as a general population measure.

An argument against a causal link between low cholesterol and disease is the fact that the levels of cholesterol and diseases vary among different populations. The relationship is not consistent (Marmot, 1994). If mortality from other causes increases with cholesterol lowering then intervention would be of doubtful value. Falling cholesterol concentrations could have different biological effects than consistently low concentrations. If cholesterol lowering itself is not harmful, and diet is safe, then attention must focus on the side effects of particular drugs. Drugs will be appropriate for those at higher risk only when the benefits outweigh the hazards.

Following on from this review of previous studies of coronary heart disease, the next chapter will consider the West of Scotland Coronary Prevention Study (WOSCOPS) in more detail. This is a randomised clinical trial of the safety and efficacy of the cholesterol lowering drug pravastatin, for the primary prevention of CHD.

Chapter 3 will then examine an alternative method of adverse event follow-up for clinical trials using record linkage techniques, and chapter 4 will consider how record linkage techniques have been applied to WOSCOPS.

Chapter 5 will compare this alternative method of follow-up using record linkage with the more traditional follow-up methods routinely applied in WOSCOPS, in order to assess the quality of the follow-up achieved by each of these methods in the context of a clinical trial.

Having thus assessed the quality of adverse event data obtained by record linkage for the clinical trial branch of WOSCOPS, chapter 6 gives an analysis of the data acquired for the WOSCOPS screened cohort by this technique, in relation to baseline risk factors. Analysis of this large cohort will address some of the issues raised in this chapter. The outcomes of all-cause mortality, coronary heart disease deaths, cancer deaths and trauma deaths will be examined, as well as cancer incidence. Analysis will consider the coronary heart disease risk factors discussed above, with primary emphasis on cholesterol, alcohol consumption and a measure of socio-economic status.

Chapter 7 will compare the event rates observed in the screened cohort (via record linkage) to the event rates for the population in the area of screening, in order to assess how representative of the general population a screened cohort is.

Finally, chapter 8 will give a review of the conclusions which can be drawn from this thesis, and suggestions for future work to be undertaken.

Chapter 2

The West of Scotland Coronary Prevention Study

2.1 Background to the study

Previous epidemiological studies have consistently supported the link between raised plasma cholesterol levels and increased rates of coronary heart disease mortality, both individually and combined (Stokes et al, 1987; Law, Wald and Thompson, 1994). The three largest of the previous clinical trials of cholesterol lowering agents, the WHO clofibrate study (Committee of Principal Investigators, 1978), the Lipid Research Clinics Coronary Primary Prevention Trial (Lipid Research Clinics Program, 1984) and the Helsinki Heart Study (Manninen et al, 1988) have also supported this causative link, although their results have been disappointing due to low statistical power. These and other studies have been discussed in Chapter 1.

The West of Scotland Coronary Prevention Study (WOSCOPS) was the fourth large clinical trial to be carried out on the use of cholesterol lowering drugs for the primary prevention of coronary heart disease. WOSCOPS was a double-blind, randomised, placebo-controlled primary prevention trial in middle-aged men (45-64 years) with raised cholesterol levels. It was designed to test the long-term safety and efficacy of the cholesterol lowering drug pravastatin, a competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A (HMG CoA) reductase, in the primary prevention of coronary heart disease morbidity and mortality. The HMG CoA reductase inhibitors have been shown to be potent reducers of LDL cholesterol with reductions of 30% being readily achievable with once per day treatment. Early experience with this class of drugs has shown good compliance and a low incidence of adverse events. Active treatment

should result in reductions of up to 34% in LDL cholesterol, as much as 25% in plasma triglyceride, and mean increases of up to 14% in HDL cholesterol, thus significantly improving the whole lipoprotein profile (La Rosa, 1989).

The design and administration of WOSCOPS are described in more detail elsewhere (West of Scotland Coronary Prevention Study Group, 1992). The major trial endpoint was coronary heart disease death plus non-fatal myocardial infarction. The study also reported on the incidence of coronary mortality, all-cause mortality, coronary artery bypass graft, angioplasty, coronary arteriography and cerebrovascular disease. 90% of the subjects lived within a 30 mile radius of the Data Centre in the University of Glasgow, with the remaining 10% coming from the area of Dumfries and Galloway approximately 90 miles to the south of the city. All randomised subjects gave informed consent including written permission for the follow-up of their medical records. Participants were monitored for an average of 5 years to obtain long-term efficacy and safety information. By the end of the study, WOSCOPS had accumulated approximately 32,000 patient years of follow-up.

The progress of the trial was reviewed regularly by an external Data and Safety Monitoring Committee, which was responsible for recommending the continuation or termination of the trial in the light of any observed treatment effects (whether pronounced adverse effects or significant benefits). This external committee was the only body with access to unblinded information during the course of the trial.

In WOSCOPS, the statistical power was improved, as compared to the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) and the Helsinki Heart Study (the 2 more recent large clinical trials of cholesterol lowering drugs), by

1) using a larger sample size, with 6595 men randomised to either pravastatin or placebo (an increase of 50% compared to the LRC-CPPT (3806 men) and the Helsinki Heart Study (4081 men)).

2) recruiting older men, age 45-64, with a corresponding higher cardiac event rate (mean age=55.2 in WOSCOPS, mean age=47.8 in LRC-CPPT and mean age=47.3 in Helsinki Heart Study).

3) recruiting men in a high risk area (the West of Scotland has among the highest rates in the world).

4) using a powerful new cholesterol reducing agent, pravastatin, which has the potential for reducing the cardiac events in the active treatment group by at least 30% over the average 5 years of follow-up. This drug is substantially more palatable than the drugs used in the previous studies, easy to take and with few apparent side effects, and should thus ensure a higher compliance with the study protocol (West of Scotland Coronary Prevention Study Group, 1992).

In addition to the clinical trial side of WOSCOPS, the population screening programme provided important epidemiological information on coronary heart disease risk factors in the West of Scotland. There has been substantial scientific and health board interest in following these people up, to study relationships between baseline risk factors and incidence of heart disease, cancer and other illnesses.

The process of recruitment involved the initial population screening programme followed by two filtering visits and a randomisation visit. This led to different levels of data being available for subjects who reached each stage.

2.2 Screening visit 1

In order to randomise 6595 men into the study, it was necessary to carry out a large population screening exercise, which ran from October 1988 until March 1991. Initially, screening was open to all adults, but as time went on, more emphasis was placed on screening men in the targeted age range of 45-64, who were identified from doctor's age-sex registers or local health centre computerised databases (GPASS), and invited by mail to come for screening. Screening took place in four health board areas, Greater Glasgow Health Board, Lanarkshire Health Board, Argyll and Clyde Health Board, and Dumfries and Galloway Health Board, all of whom gave ethical permission for the study.

Of the 105,383 subjects who attended an initial screening visit, 16776 were women, and 88,607 were men, of whom 81,161 were age eligible. At this first screening visit,

subject identifying information was collected, as well as information on risk parameters, personal history, clinical findings and demographic variables. While baseline measurements of risk factors are available for all of these 105,383 subjects, only 97,165 subjects had complete name, address and date of birth information recorded.

The full list of variables measured can be seen in Table 2.1. At this first screening visit, data were entered directly into a computer database by the screening centre nurses and no case report forms were completed.

Identification	Name Address Date of birth Sex Marital status (see below) Date of screening visit 1
Demography	Educational level (see below) Occupational category (see below)
Risk parameters	Height (cms) Weight (kg) Aerobic exercise (hours per week) Diastolic blood pressure (mmHg) Systolic blood pressure (mmHg) Cholesterol (mmol/l) Smoking habit (see below) Number of cigarettes smoked per day Number of years smoked Number of years stopped smoking Alcohol intake (units per week)
History / clinical findings	Personal history of CHD (yes/no) Personal history of diabetes (yes/no) Personal history of hyperlipidaemia (yes/no) Personal history of hypertension (yes/no) Number of first degree relatives dead from CHD Presence of xanthomata (yes/no) Presence of corneal arcus (yes/no)

Table 2.1 Data recorded at screening visit 1

The categorical variables are classified as:-

Marital status:	single	Education:	left school < 16
	married		highers
	widowed		technical college
	divorced		university
	separated		other
Occupation:	unemployed	Smoking:	never smoked
	unskilled		ex-smoker
	skilled		current cigarette smoker
	managerial		current pipe / cigar smoker
	professional		
	other		

At this stage, the cholesterol measurement related to Total Cholesterol as measured using a Reflotron bench-top analyser. The obvious physical signs of xanthomata and corneal arcus, possibly indicating hypercholesterolaemia, were also noted at this visit.

Two derived variables were calculated from this recorded information:

Age = Date of screening visit 1 - Date of birth, and Body mass index = $\frac{\text{weight}}{\text{height}^2}$

Men in the targeted age range who had a total cholesterol level $\geq$ 6.5 mmol/l, no knowledge of a previous myocardial infarction and who were willing to participate were given dietary advice on cholesterol reduction and invited to return for a further screening visit in 4 weeks time, as a potential recruit.

2.3 Screening visit 2

From the second screening visit onwards, all data collected at the screening centres were recorded on standard case report forms. Approximately 20,800 men with high cholesterol returned for a second screening visit, when they underwent a fasting lipoprotein analysis, in which total plasma cholesterol, triglyceride, very low density

lipoprotein (VLDL), low density lipoprotein (LDL) and high density lipoprotein (HDL) cholesterol were determined. If the LDL cholesterol level was ≥ 4.0 mmol/l then the subject was asked to return in another 4 weeks for a further screening visit. Blood samples were sent to Glasgow Royal Infirmary for analysis accompanied by a form containing the patient's name, address, GP information, centre number and subject number (this form, containing patient identifying information, is given in Appendix A). All subsequent screening forms identified the patient only by a patient number and the patient's initials. At this second screening visit, patients were also asked about their past and current general health and family history, with particular emphasis on coronary events, and the Rose Questionnaire for detection of angina pectoris and intermittent claudication was administered.

2.4 Screening visit 3

Approximately 13,600 subjects with moderately raised LDL cholesterol and no reported history of CHD returned for a third screening visit at which a further fasting sample was taken for lipoprotein analysis, together with blood for a full biochemical and haematological profile and measurement of fibrinogen and plasma viscosity. The biochemical and haematological variables measured are given in Table 2.2. A 12 lead resting ECG was performed at this visit. More detailed information on the subject's current health, alcohol consumption, smoking habits (past and present), and demographic details such as education and employment status were also recorded.

Biochemical Tests	Haematological Tests
AST (u/l)	MCV (fl)
ALT (u/l)	Haemoglobin (g/l)
ALP (u/l)	White cell count (cell/l)
CK (u/l)	Red cell count (cell/l)
Calcium (mmol/l)	
Total protein (g/l)	
Creatinine (umol/l)	
Glucose (mmol/l)	
Sodium (mmol/l)	
Potassium (mmol/l)	
Bilirubin (mmol/l)	
Triglyceride (mmol/l)	

Table 2.2 Biochemical and Haematological Tests carried out at Visit 3

The main inclusion criteria for the trial were based on LDL cholesterol levels as follows:

- LDL cholesterol $\geq$ 4.0 mmol/l (154 mg/dl) at both screening visits 2 and 3
- LDL cholesterol $\geq$ 4.5 mmol/l (174 mg/dl) at either screening visits 2 or 3
- LDL cholesterol $\leq$ 6.0 mmol/l at either screening visits 2 and 3

This requirement of a raised LDL cholesterol level on 2 occasions, measured several weeks apart, meant that subjects whose cholesterol level became acceptable simply by following the dietary advice given were not randomised to study treatment. These requirements also excluded subjects who consistently showed a very high LDL cholesterol level possibly requiring overt therapy.

Patients with any evidence of previous myocardial infarction (including evidence from their screening visit 3 ECG), other life threatening illnesses or any physical or mental disability which might interfere with completion of the study, were excluded, as were patients with hypertension despite treatment or who were already on lipid lowering therapy. Subjects with a positive Rose Questionnaire were only excluded if they had been hospitalised for treatment or investigation of angina within the previous 12 months. There were further exclusion criteria based on the biochemical and haematological tests carried out at screening visit 3 (West of Scotland Coronary

Prevention Study Group, 1992). Biochemical tests provided a general screening to try to exclude patients with possible concurrent disease. The tests of liver function (in particular AST and ALT (serum transaminases)) and the muscle enzyme creatine kinase were carefully monitored throughout the study for safety reasons, so a baseline level within normal limits had to be established.

2.5 Screening visit 4

The above screening procedures resulted in 6595 subjects being randomised to treatment with either placebo or pravastatin at their fourth screening visit. These randomisation visits took place between February 1989 and September 1991.

At the randomisation visit (visit 4), a trial physician conducted a general physical examination and an opthalmoscopic examination for lens opacities and visual acuity, and further details were collected on blood pressure, heart rate, height, weight and concurrent medication. If the subject was willing to take part in the study, then written, informed consent was obtained and the subject was randomised. All screening case report forms were then transmitted to the Data Centre where the data was entered, verified and stored in a central computer database.

The baseline characteristics of the 6595 randomised subjects as recorded at screening visit 1 are presented in Table 2.3, along with the corresponding data for the screened population of men in the same age range.

	All Screenees	Randomised Subjects
Continuous variables		
Age at visit 1 (years)	54.8 (5.7)	54.8 (5.5)
Systolic BP (mmHg)	137.4 (19.2)	138.2 (18.5)
Diastolic BP (mmHg)	84.9 (10.8)	85.4 (10.5)
Weight (kg)	77.5 (12.4)	78.5 (11.0)
Body mass index (kg/m^2)	25.8 (3.7)	26.1 (3.2)
Total cholesterol (mmol/l)	5.9 (1.2)	7.3 (0.7)
Categorical variables		
History of hypertension	16.1	14.7
History of hyperlipidaemia	2.6	4.0
History of CHD	11.6	4.6
Current cigarette smoker	36.5	36.7
Alcohol (units per week)		
<21 units	80.4	82.0
>= 21 units	19.6	17.8

Table 2.3 Comparison of data recorded for all age eligible men (81,161 subjects) and the randomised subjects (6595 subjects). Continuous variables are presented as mean (standard deviation). Data for categorical variables are given as percentages of all subjects in the 'All Screenees' group or 'Randomised Subjects' group as appropriate.

The above data suggest that the screening process has not resulted in the selection of a particular subgroup which is substantially different from the screened population other than with respect to cholesterol level and previous history of coronary heart disease (West of Scotland Coronary Prevention Study Group, 1995).

2.6 Subject follow-up during the trial

Patients were followed up in the community, and attended trial centres set up either in GP's surgeries, local health centres or other public buildings at 3 monthly intervals, for monitoring of adverse events (including potential side-effects) or endpoint events, recording of concurrent medication, re-testing of lipid levels and monitoring of clinical

and laboratory safety parameters, as well as receiving a new supply of tablets and returning unused tablets to be counted for compliance assessment. The data on the lipid levels (post-baseline) were not available to the investigators, in order to maintain blinding, and were used only for safety purposes. The information collected at each visit was as in Table 2.4 with patients being required to fast overnight prior to 6 monthly and annual visits. Participants were also given regular dietary advice and actively encouraged to stop smoking.

Procedure	**3 / 9 monthly visits**	**6 monthly visits**	**12 monthly visits**
Compliance pill count	x	x	x
Concurrent medication	x	x	x
Adverse event check list	x	x	x
Record hospitalisations or other endpoints	x	x	x
Intercurrent illness	x	x	x
Physical examination			x
BP, smoking record		x	x
Full lipoprotein profile		x	x
Biochemical profile		x	x
Haematology			x
ECG			x
Angina questionnaire			x
Opthalmic tests			x

Table 2.4 Procedures performed at follow-up visits

At 3, 6 and 9 monthly visits, the subject was seen by a nurse and the documentation checked by a trial physician, while annual (12 monthly) visits were conducted by the trial physician. The trial physician may have interviewed the patient at any visit or at times between visits however, if they reported an adverse event. Opthalmic tests were carried out on an annual basis to check for visual acuity and lens opacity (initially thought to be possible side-effects of the cholesterol lowering therapy). Any patient displaying lens opacity was followed up closely by an opthalmologist. At the first annual visit, a second measurement of fibrinogen and plasma viscosity was made.

All data entry and validation of case report forms was carried out at the Robertson Centre for Biostatistics at Glasgow University and central biochemical and haematological analyses of blood samples took place at Glasgow Royal Infirmary. Automated ECG analysis was also undertaken at Glasgow Royal Infirmary using the Glasgow program (MacFarlane, Devine et al, 1990) to generate Minnesota codes, a scheme for classifying ECG abnormalities (MacFarlane, Latif et al, 1990).

Documentation of adverse events relied primarily on patients' self-reporting. WOSCOPS personnel had no direct access to a patient's medical records even when trial visits took place at the subject's own health centre. This process was supplemented through manual flagging with the National Health Service Central Register (NHSCR) for deaths and incident cancers.

Information on adverse events was recorded by the medical staff based at each trial centre, and reviewed by an Adverse Events Committee. Support documentation was obtained by a clinical researcher from Glasgow Royal Infirmary for all serious adverse events, including deaths, cardiac events and cancers. The Adverse Events Committee consisted of three consultant physicians who examined all serious or cardiac adverse events and allocated an ICD code to them. They also indicated whether an event should be passed to the Endpoints Committee or not. Possible study endpoints were classified by the Endpoints Committee, who reviewed all deaths, possible myocardial infarctions and annual study ECGs which showed serial changes. The Endpoints Committee was staffed by two cardiologists and an ECG specialist who decided whether or not a given adverse event met the criteria of an endpoint. Both the Adverse Events Committee and the Endpoints Committee remained blinded throughout the study. Safety monitoring of abnormal bloods and general adverse events was carried out at Glasgow Royal Infirmary by Clinical Co-ordinators.

Subjects who withdrew from taking study medication were encouraged to attend for an annual visit so that information on adverse events could be obtained, along with a blood sample and ECG. This was important, as final analysis was carried out on an intention-to-treat basis, with an outcome being sought for every patient.

Adverse events were all documented on an adverse event report form. Serious adverse events, however, triggered specific additional forms depending on the nature of the event. More than one additional form could be completed for events as required. The documentation structure is illustrated in Figure 2.1 and examples of some of these forms are given in Appendix B. Separate forms were completed for a hospitalisation, death, possible myocardial infarction, coronary arteriography, angioplasty, coronary artery bypass graft, cerebrovascular accident (stroke) or cancer. A form confirming the event to be a trial endpoint was also completed by the Endpoints Committee where appropriate.

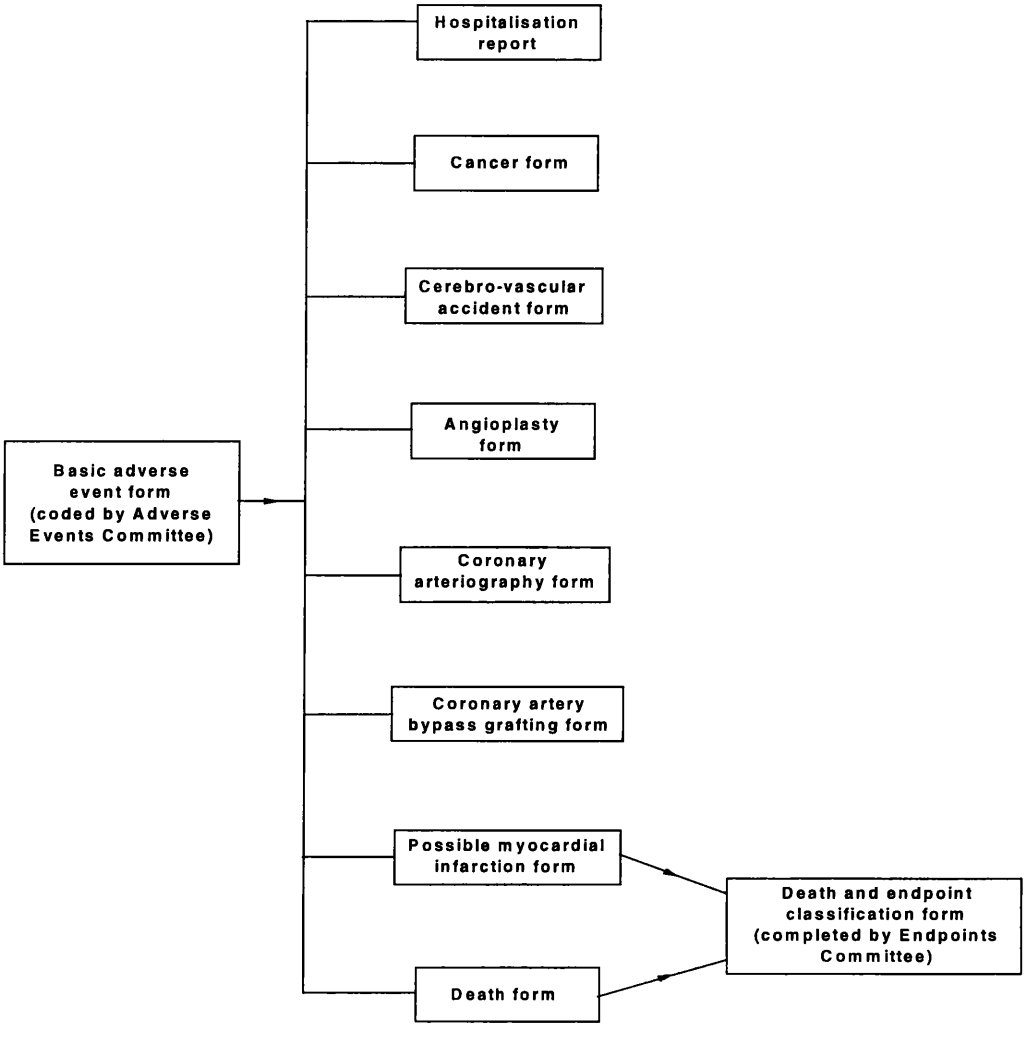


Figure 2.1 Structure of Adverse Event Documentation

Follow-up was completed at visits scheduled between January and May 1995, at which a final study ECG was recorded. The vital status of all subjects (including those who failed to attend their visit) was ascertained during this period.

2.7 Discussion of the Adverse Event Reporting System

WOSCOPS adverse event follow-up by patient self-reporting has been supplemented by flagging the WOSCOPS subjects with NHSCR. Since 1948, NHSCR has contained brief particulars of every person registered with a GP, updated for deaths and emigration where known. In the 1970s, information from the National Cancer Register was added, providing a system by which study subjects can be flagged for all deaths and incident cancers, by comparing the identifying particulars of study subjects with those known to NHSCR. Currently, this matching is done clerically, although moves are being made to automate this system (Acheson, 1987).

A study was carried out in Oxford to assess the efficiency of NHSCR notification of breast cancers which had been verified as part of a cohort study of women taking HRT (Hunt and Coleman, 1987). 28 out of the 50 cases observed in the study had not been notified by NHSCR two and a half years after the diagnosis of the most recent case. Of these 28 cases, 14 had not been registered (a problem arising from the fact that in the UK, cancer registration is voluntary), 8 had been registered but had not yet arrived at NHSCR, 5 were in the process of being notified, and 1 registration could not be linked to the individual's record. It was not uncommon for there to be a year between receipt of a batch of registrations at the central registry and processing at NHSCR, on top of the delay between diagnosis and registration (median delay was 1 year). This study took place in England, where 12 cancer registries send data to the Office of Population Censuses and Surveys (OPCS) centrally, but there is no reason to expect that the situation is any better in Scotland, where 5 registries send data to the Information and Statistics Division (ISD) of the Scottish Health Services Common Services Agency for central processing. Notification of deaths is regarded as being reliable and complete, but there is concern about the completeness and timeliness of cancer notifications.

Concerns over the completeness of adverse event information derived from patient reporting and NHSCR, have led to the consideration of an alternative method of patient follow-up using record linkage techniques.

Scotland is fortunate in having national databases covering all deaths, incident cancers, hospitalisations, psychiatric hospitalisations and cardiac surgical procedures. These databases are all held in Edinburgh within the Scottish Record Linkage System (Kendrick and Clarke, 1993). Record linkage techniques, which will be described in chapter 3, made it possible to obtain details of adverse events for the WOSCOPS subjects from these national databases, as described in chapter 4, using the patient identifying information recorded at screening visits 1 and 2 (see Appendix A). The more traditional methods of subject follow-up for clinical trials could thus be validated and supplemented, as discussed in chapter 5. Record linkage also provided a cost effective manner in which prospective epidemiological follow-up could be carried out for the large cohort of subjects who attended screening visit 1.

Chapter 3

Record Linkage Methodology

3.1 What is record linkage?

'Record Linkage' is simply the bringing together of information from two records believed to relate to the same individual. When this is based on the calculation of the likelihood of a correct linkage (that is, that the individuals represented on the two records are really the same person), the process is said to be 'probabilistic'. Probabilistic record linkage is appropriate where the identifying particulars used for searching and 'linking' records are prone to change or are less than reliably reported, or when different subjects may have the same identifiers. Record linkage techniques are used subconsciously by everyone on a daily basis, for example, when looking up a number in a phone book. In this example, the name of the person whose phone number is required is compared to a 'file' of names until a 'match' is found. Record matching algorithms are at the heart of all automatic information retrieval systems. The use of record linkage systems in the context of clinical trials and cohort studies will be described in Chapter 4.

Although there are a few early references to the problems of record matching (Farr, 1861; Stocks, 1944), the main stimulus to research in this area came in the 1960s with attempts to store medical and vital status records on a computer system. As ever more data are stored in this way, the scope for record matching and, in particular, matching with a probabilistic element increases correspondingly. The statistician has a role to play in the production of matching algorithms, assessment of evidence for matching, and evaluation of the accuracy of results from statistical matching, where there is any uncertainty over the choice of matched pairs.

3.2 Standard Methodology for Record Linkage

3.2.1 Empirical concepts

Much of the early development of record linkage techniques was carried out by Howard Newcombe and colleagues in Canada (Newcombe et al, 1959). He describes the probabilistic basis of record linkage and a practical plan for setting up a record linkage system in his book, the Handbook of Record Linkage (Newcombe, 1988). This section gives Newcombe's suggestions on how record linkage may be carried out in practice.

The degree of certainty of a correct linkage obviously depends on the outcomes from the comparisons of individual identifiers. If all the identifiers agree and are unlikely to have done so by accident, then the level of assurance of a correct link will be high, but if they all disagree and are unlikely to have done so in records truly belonging to the same person, then there is little doubt that the pair are wrongly matched. Partial similarities and dissimilarities may argue in either direction. Probabilistic record linkage quantifies this comparison process and decides where the balance lies for intermediate situations where some of the identifying information points to a link and other information does not. Often an identifier may fail to agree precisely on a pair of records but be obviously similar nevertheless. The levels of agreement and disagreement can be subdivided into appropriate categories of partial match/mismatch, and frequency ratios calculated accordingly.

The basic question for each comparison is 'How typical is that comparison outcome among pairs of records relating to the same person ("linked"), as compared with pairs of records brought together at random ("unlinkable")?' The first step to addressing this question is the calculation of frequency ratios. The **frequency ratio** represents the 'betting odds' in favour of a correct match, associated with the particular comparison and its outcome.

$$\text{Frequency Ratio} = \frac{\text{frequency of outcome (x, y) among linked pairs}}{\text{frequency of outcome (x, y) among unlinkable pairs}}$$

x = value of identifier on file initiating search

y = value of identifier on file being searched

The combined frequency ratios pertaining to a given record pair (that is, their product) constitute the information on which the overall 'betting odds' in favour of a correct match are based.

In order to calculate these frequency ratios, one first needs to manually produce a file of a few hundred correctly matched pairs of records, in order to calculate the frequencies of the various comparison outcomes in this file of linked pairs. A larger file of record pairs brought together randomly will provide the frequencies of the outcomes for unlinkable pairs.

The simplest comparisons use 'global' discriminating powers where the frequency ratios are derived from the files of linked and unlinkable pairs, but are non-specific for the value of an identifier. For example, where the surname is the same on both records, but no account is taken of the rarity of that surname in the population under consideration. Thus, global treatments under-exploit the discriminating power of the information available.

In practice, global frequency ratios are routinely converted to their 'value-specific' counterparts wherever this would result in better use being made of the discriminating power of the identifiers. The frequencies of specific values for such identifiers may be stored in look-up tables.

$$\text{Value-specific freq ratio} = \frac{\text{freq of agreement in linked pairs (A has specified value)}}{\text{freq of agreement in unlinkable pairs (A has specified value)}}$$

where record A is from the file that initiates the search.

Both the numerator and the denominator of the value-specific ratio are usually estimated rather than observed. For disagreements, it is assumed that the particular value of the identifier on record A does not influence the chance of a disagreement. The global odds for a given disagreement now provides an estimate for the value-specific odds for the disagreement. For full agreements, it is assumed that in the linked pairs, the particular value of the identifier on record A does not influence the chance of an agreement. The global numerator for a given agreement thus provides an estimate for the value-specific numerator. It is also assumed that the frequency, in the unlinkable pairs, of a chance agreement of the particular value of the identifier on record A will be identical to the frequency of that particular value in the file being searched, thus providing an estimate of the value-specific denominator for full agreement. However, this simple approach is inappropriate for partial agreements, due to difficulties calculating the value-specific denominator.

While the ideal method of obtaining the denominator of these outcome frequencies is from an actual file of unlinkable pairs, in practice it is often possible to calculate approximations to them based on various assumptions. For example,

a) Comparison of month and day of birth - Births are assumed to be uniformly distributed over the months of the year, and over the days of the month. The general formula to obtain the frequencies of various outcome levels, when the records are matched randomly is

$$\text{Frequency of discrepancy } x = \frac{n - x}{n^2}$$

where n is the number of months in a year, or days in a month and x is the magnitude of the positive or negative difference between the two months, or days. If both positive and negative differences are to be included together, then (n-x) must be multiplied by 2. Slight seasonal variation in the birth pattern has little effect on the accuracy of this calculation, which yields outcome frequencies almost exactly the same as those found in an actual file of unlinkable pairs.

b) Comparison of year of birth - For years of birth, it is not reasonable to assume a uniform distribution. However, for a particular calendar year, the frequency with which there will be an exact agreement by pure chance, for randomly matched pairs, will be equal to the

frequency of that year in the search file times the corresponding frequency in the file being searched. Where the year itself is unimportant, a value non-specific frequency over all years of birth can be found by summing the frequencies for individual years. Levels of disagreement can be dealt with in a similar manner.

A simplified, empirical approach to calculating value-specific frequency ratios is to convert the global ratio to it's value-specific counterpart using an adjustment factor, rather than calculating a new ratio. The adjustment factor will depend on whether the agreement portion of the specific identifier is rare or common in comparison with the general frequency for that portion of the identifier. The general frequency for an identifier is a weighted mean of all the specific frequencies in the file being searched.

$$\text{General frequency} = \sum_{x} (\text{frequency of value } x)^2$$

$$\text{Adjustment factor} = \frac{\text{general frequency}}{\text{value - specific frequency}} \qquad \text{in file being searched.}$$

For convenience, the ratios of the two frequencies are often expressed logarithmically and called weights. Value-specific weights (or adjustment weights) are usually stored in look-up tables, while non-specific global weights are written directly into the comparison rules.

It is often simpler to regard a missing comparison as indicating nothing. But the assumption that blank fields are essentially neutral may not be true. For example, a blank middle initial may indicate that the subject has no middle name, rather than that an existing middle name has been missed out. If blanks could be significant, they should be treated as another specific value option when calculating frequency ratios.

Identifiers that are logically related may appear on records as separate identifiers or as a single identifier (for example, first initial plus remainder of first name, or home postcode plus health board of treatment). The two parts may be compared separately, with the second comparison being conditional on the outcome of the first, or they may be combined and concatenated, with recognition of a number of possible comparison outcomes. But

care must be taken in the calculation of frequency ratios for these conditional comparisons, and they should not be considered as separate, independent comparisons.

Finding appropriate discriminating rules is a trial-and-error process, requiring repeated refinements, with the ultimate purpose being for the computer to adopt the strategies of the human mind in order to keep expensive clerical resolutions to a minimum. More sophistication in the comparison procedures, possibly involving unconventional comparisons or more levels of similarity, are necessary if the numbers of false-positive and missed linkages are important, and manual checking is to be kept to a minimum. Much time must initially be spent resolving borderline links, to get clues as to how they are resolved, and thus develop the comparisons. However, there is a point beyond which the cost of refining rules outweighs the advantages of applying them, particularly if the refinement requires an extensive amount of manual review.

The overall odds in favour of a correct link is found by multiplying together all of the frequency ratios for the comparison pair. This is usually done by converting each of the frequency ratios into its logarithm, called a weight, and adding the weights to produce a total weight. This assumes that the various agreements and disagreements are independent of each other. For a single-step multiple-outcome procedure, the combined frequencies of all outcomes should add to 100%, among both the linked and unlinkable pairs. Total weight is a relative measure of the assurance of a correct link, not an absolute measure. It chiefly serves to rank the matched pairs in order of assurance, but does not indicate what the actual 'betting odds' would be. A threshold weight at which the records will be linked is set by clerical checking. Strictly speaking, total weights reflect only the likelihood or unlikelihood that the observed similarity of identifying information has arisen other than by chance. But the ruling out of chance does not necessarily establish that the same person is involved (Newcombe et al, 1983). For example, the similarity could be due to twins.

The accumulated frequency ratios take no account of the probability that a search record is in the file being searched, or the size of that file. For example, you might expect only 10% of your study population to have died, and thus appear in a death file. When relative odds only are used, a subjective threshold is set in the transition region from good to bad links. It would also be possible to subjectively set two thresholds to separate good, doubtful and

bad links. The problem of establishing thresholds is a function of the far greater size of the nonlinks relative to the links.

, the more objective method of assessing the absolute odds (in Newcombe's terminology) has distinct advantages, such as making it easier to detect unexpected sources of bias, and giving the true 'betting odds' that the record pair is correctly matched. Using absolute odds rescales the odds, and thus brings the subjective threshold closer to the odds point required (for example, 50:50).
The conversion formula is:

$$\text{AbsoluteOdds}=\text{RelativeOdds}\times\frac{\#\text{ linked search records}}{\text{total }\#\text{ search records}}\times\frac{1}{\text{total }\#\text{ records being searched}}$$

where the search records are from the file initiating the search (for example, study subjects) and the records being searched are from the master file of information (for example, Registrar General Death records).

So the absolute odds in favour of a correct linkage, and its logarithm the absolute total weight, are determined by the prior probability that a search record will find a correctly matching record in a single random draw from the file being searched (made up of the probability that a correct match exists, and the number of records in the file) and reflect the relative odds of the combined frequency ratios. A further possible adjustment would be to make the absolute odds age-specific. For example, considering death records, a young person would be far less likely to appear in a death file than an old person, so year of birth agreement for a young person would carry more discriminating power than for an old person, in a given year.

When relative odds have been converted to absolute odds, the single optimum threshold (minimising linkage errors) is likely to be near the 50:50 odds point, although a different balance of false positive to false negative may be desirable. Thus, a conscious decision must be taken concerning the desired ratio of the two types of error (false match or missed link). However, a 50:50 absolute odds ratio serves only as a rough guide to placing an optimum threshold.

The primary personal identifiers for linkage are sex, names, date of birth and geographical information. To avoid the problem of different spellings of surname, a phonetic code is used. There are two common phonetic coding systems, the New York State Intelligence Information System (NYSIIS) and the Soundex, of which the NYSIIS is considered to have more discriminating power. These codes suppress vowel information (with Soundex losing vowels completely, while NYSIIS at least retains a marker of where the vowels were) and replace certain consonants by a standard character representing that sound. Both the phonetically coded version and the alphabetical name should be compared, with the alphabetical comparison being conditional on an agreement on the phonetic code. An 'ill-spelled name routine' will detect levels of similarity between phonetic codes which are not exactly the same (for example, if there has been an insertion or deletion in a name), but the benefits of this must be balanced against the increased complication in calculating an appropriate frequency ratio. Cross-comparisons should also be used for initials as people may reverse the order of their given names. Adjustment factors should be applied for initials that agree so that rare initials raise the frequency ratio by more than common initials. If there is reason to suspect that they could have been inverted, cross-comparisons could also be used for day and month of birth.

The comparison of given names was considered in more detail in a recent paper (Newcombe, Fair and Lalonde, 1992). Not all variations on given names are necessarily truncations which can be recognised simply by the computer. A human checker can recognise nicknames, ethnic variants, diminutives or misspellings, for example, Anthony - Tony, Joseph - Joe, William - Bill. If a machine is to acquire a similar ability, it must learn from past experience, in the same way as humans do. Some kind of grouping of possible synonyms is inevitable, but this must be exceedingly flexible if discriminating power is not to be wasted. Informal versions of names are more likely to be used on 'alive' records than on death records. For example, Statistics Canada has created a composite file of 64,937 linked pairs of male given names derived from 26 different linkage projects involving the Canadian mortality database which gives the frequency of combinations of names which were correctly linked. The chief problem with this file is the high number of possible value pairs which are rare, or as yet unobserved in the available links. Grouping is necessary to bring rare synonyms into the same group as common forms, an unavoidably

subjective process except that variants yielding widely different odds on their own should not be put into the same group.

There are various possible shortcuts to calculating the odds in favour of a match (each with corresponding refinements). For example, pooling of first and second given names - reducing the number of look-up tables for value-specific frequencies, recognising the specificities of only the agreement portion of names which partially agree, or pooling complementary partial agreements. However, these shortcuts result in slightly increased error rates (to extents which vary according to the specific values of the given names), so their merit must be balanced against the time and effort which they save. Only when the full specificities are taken into account does the discriminating power get efficiently exploited.

The emphasis in this paper (Newcombe, Fair, Lalonde, 1992) differs from that of procedures based on degrees of phonetic similarity plus lists of exceptions, in that both similarities and dissimilarities are recognised, and necessary data are drawn from large accumulations of linked pairs of records. Archiving empirical data from past linkage studies allows comparison of the performances of different systems, and facilitates semiautomated 'learning' from past experience. The approach described here follows a general trend in statistics to develop empirical reference distributions using computers, rather then rely mainly on theoretical distributions. The complexity of the procedures involved need not be a barrier, since once they have been developed, they can be used repeatedly for many different linkage jobs. However, Arellano warns (Arellano, 1992) that the data accumulated from previous linkages must represent the same population as the new linkage, for the 'past experience' to be of relevance. He also warned that routine cross-comparisons can be very wasteful of resources if the nature of the data does not call for them, since in most linkage evaluations, 85-90 per cent of correct linkages have exact agreement on name and birthdate. Before one can learn from past experience, there must also be a rigorous definition of a successful linkage exercise.

3.2.2 Formal theory

While Newcombe and others at Statistics Canada focused on the more practical aspects of setting up a record linkage system, Fellegi and Sunter developed a mathematical theory to provide a framework for a computer-oriented solution to the problem of recognising those records in 2 separate files which represent identical persons, objects or events (Fellegi and Sunter, 1969). The theory leads to a linkage rule which is similar to Newcombe's intuitive approach.

Notation

There are 2 populations, A and B, giving two files to be linked, L_A and L_B. Let a and b denote arbitrary elements of A and B respectively. It is assumed that some elements will be common to both populations.

The set of ordered pairs $A \times B = \{(a,b); a \in A, b \in B\}$ is the union of the matched pairs $M = \{(a,b); a = b, a \in A, b \in B\}$ and the unmatched pairs $U = \{(a,b); a \neq b, a \in A, b \in B\}$.

The files for comparison, L_A and L_B, are produced by applying the record generating process to random samples from populations A and B. This process introduces some errors and incompleteness into the records, for example, from misrecording or data entry error.

The first step in attempting to link the records of the two files, L_A and L_B, is the comparison of the records, which results in a set of comparison outcomes, consisting of an outcome for each individual comparison made between the two files.

The comparison vector is: $\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \ldots\ldots, \gamma^k[\alpha(a), \beta(b)]\}$ where $\alpha(a)$ is the record for element a, $\beta(b)$ is the record for element b, and the γ^i may take the values 'agree' or 'disagree'. For example, γ^1 could compare surname and γ^2 could compare year of birth. The set of all possible realisations of γ is called the comparison space, and denoted Γ.

In the course of linkage we observe $\gamma(a,b)$ and decide that either:

- (a,b) is a matched pair, $(a,b)\in M$ — decision D_1 — positive link
- (a,b) is an unmatched pair, $(a,b)\in U$ — decision D_3 — positive non-link
- decision cannot be made at the specified levels of error — decision D_2 — possible link

The Linkage Rule, L, is defined as a mapping from the comparison space Γ onto a set of random decision functions $d=\{d(\gamma)\}$

$$d(\gamma)=\{P(D_1|\gamma),P(D_2|\gamma),P(D_3|\gamma)\} \quad \gamma\in\Gamma \qquad \sum_{i=1}^{3} P(D_i|\gamma)=1$$

The conditional probability of γ given that $(a,b)\in M$, denoted by $m(\gamma)$, is

$$P(\gamma[\alpha(a),\beta(b)]|(a,b)\in M)=\sum_{(a,b)\in M} P\{\gamma[\alpha(a),\beta(b)]\}.P[(a,b)|M]$$

The conditional probability of γ given that $(a,b)\in U$ is denoted as $u(\gamma)$ similarly.

The first type of error occurs when an unmatched comparison is linked (false match), and has probability

$$P(D_1|U)=\sum_{\gamma\in\Gamma} u(\gamma).P(D_1|\gamma)$$

The second type of error occurs when a matched comparison is not linked (missed link), and has probability

$$P(D_3|M)=\sum_{\gamma\in\Gamma} m(\gamma).P(D_3|\gamma)$$

A linkage rule on the space Γ is said to be a linkage rule at the (error) levels μ,λ $(0<\mu<1,0<\lambda<1)$, denoted by $L(\mu,\lambda,\Gamma)$ if $P(D_1|U)=\mu$ and $P(D_3|M)=\lambda$.

The linkage rule $L(\mu,\lambda,\Gamma)$ is the optimal linkage rule if the relation $P(D_2|L)\leq P(D_2|L')$ holds for every $L'(\mu,\lambda,\Gamma)$. This optimal linkage rule minimises the probability of failing to make a positive decision.

Main Theorem

A linkage rule L_0, defined on Γ with the ordered set $\{\gamma\}$ indexed by subscript i ($u_i=u(\gamma_i), m_i=m(\gamma_i)$) is the best linkage rule on Γ at the levels (μ,λ) if $L_0(\mu,\lambda,\Gamma)$ is defined as:

Take action D_1 if $i \leq n-1$

Take action D_2 if $n < i \leq n'-1$

Take action D_3 if $i \geq n'+1$

Take a random decision if $i=n$ or $i=n'$

$1<n<n'<N_\Gamma$, ensuring that (μ,λ) are an admissible pair of errors, and n, n′ are chosen such that

$$\sum_{i-1}^{n-1} u_i < \mu \leq \sum_{i=1}^{n} u_i \qquad\qquad \sum_{i=n`}^{N_\Gamma} m_i \geq \lambda > \sum_{i=n`+1}^{N_\Gamma} m_i$$

$$d(\gamma_i) = \begin{cases} (1,0,0) \\ (P_\mu, 1-P_\mu, 0) \\ (0,1,0) \\ (0,1-P_\lambda, P_\lambda) \\ (0,0,1) \end{cases} \qquad \text{when} \quad \begin{cases} i \leq n-1 \\ i = n \\ n < i \leq n`-1 \\ i = n` \\ i \geq n`+1 \end{cases}$$

where $u_n.P_\mu = \mu - \sum_{i=1}^{n-1} u_i$ and $m_{n'}.P_\lambda = \lambda - \sum_{i=n`+1}^{N_\Gamma} m_i$

If $\mu = \sum_{i=1}^{n} u_i$, $\lambda = \sum_{i=n}^{N_\Gamma} m_i$ and we define the threshold levels $T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$ and $T_\lambda = \frac{m(\gamma_{n`})}{u(\gamma_{n`})}$ then this leads to

$$d(\gamma) = \begin{cases} (1,0,0) \\ (0,1,0) \\ (0,0,1) \end{cases} \quad \text{if} \quad \begin{cases} T_\mu \leq m(\gamma)/u(\gamma) \\ T_\lambda \leq m(\gamma)/u(\gamma) \leq T_\mu \\ m(\gamma)/u(\gamma) \leq T_\lambda \end{cases}$$

The use of threshold values T_μ and T_λ makes it unnecessary to explicitly order the values of γ in order to make decisions. Simply compare $m(\gamma)/u(\gamma)$ to the threshold values. Since, in practical situations, the number of total configurations is usually very large, T_μ and T_λ are estimated from a random sample of configurations.

This sample (size s) may be ordered by decreasing values of $w(\gamma)=\log(m(\gamma))-\log(u(\gamma))$, and the h^{th} member of the ordered listing denoted by γ_h. Then $P(w(\gamma)<w(\gamma_h)|\gamma\in M)$ is estimated by $\lambda_h=\sum_{h`=h}^{s} m(\gamma_{h`})/\Pi(\gamma_{h`})$ where $\Pi(\gamma_h)=s/2 \times z`(\gamma_h)$ and $z`$ are the probabilities that element are selected for the random sample and $P`(w(\gamma)<w(\gamma_h)|\gamma\in U)$ is estimated by $\mu_h=\sum_{h`=1}^{h} \frac{u(\gamma_{h`})}{\Pi(\gamma_{h`})}$. The threshold values $T(\lambda_{h`})$ and $T(\mu_{h`})$ are simply the weights $w(\gamma_{h`})$ and $w(\gamma_h)$.

If error levels sufficiently high to preclude action D_2 can be tolerated then n and n` can be chosen so that the middle set of γ is empty. Every record pair (a,b) will then be allocated to either M or U. So allocation of observations to one of two mutually exclusive populations by setting a single threshold level is thus a special case of the above theory.

The theory assumes that all possible record comparisons will be attempted, but in practice, to make the numbers feasible, comparisons will be made only within corresponding blocks. If the number of comparisons examined explicitly is thus restricted to a subspace Γ^* then the probabilities in the theory above should be replaced with proportions. The subspace Γ^* is then the set of γ for which the blocking component has the agreement status, and all other γ are implicit positive non-links.

In choosing the error levels (μ,λ) we may want to be guided by consideration of the losses incurred by the different actions.

$G_M(D_i)$ and $G_U(D_i)$ are non-negative loss functions which give the loss associated with decision D_i. Set $G_M(D_1)=G_U(D_3)=0$.

We seek the subspace Γ^* which minimises the total expected loss (when using a blocking procedure)

$$= c\left\{P(M)\left[P^*(D_2|M)G_M(D_2)+\lambda^* G_M(D_3)\right]+P(U)\left[\mu^* G_U(D_1)+P^*(D_2|U)G_U(D_2)\right]\right\}$$
$$+ G_{\Gamma^*}(L_A \times L_B)$$

where c is the number of comparisons in $L_A \times L_B$.

If the processing cost of comparison under any blocking Γ^* is proportional to the number of comparisons c^*, that is $G_{\Gamma^*}(L_A \times L_B) = \alpha c^*$, then we can minimise

$$P(M)\left[P^*(D_2|M)G_M(D_2)\lambda^* G_M(D_3)\right]+P(U)\left[\mu^* G_U(D_1)+P^*(D_2|U)G_U(D_2)\right]+\frac{\alpha c^*}{c}.$$

No explicit solution to this is possible under general conditions, but it can be used to compare two different choices of Γ^*. Once Γ^* has been chosen, the 'theoretical' error levels μ,λ can be chosen so that the actual levels μ^*, λ^* meet the error specifications.

The basic Fellegi-Sunter model assumes conditional independence of comparisons, in order to allow estimation of the probabilities. Fellegi and Sunter believed that their model was reasonably robust to departures from independence, but methods have now been developed to adjust for lack of independence, as described elsewhere (Winkler, 1989).

Recent advances in methodology include the use of an EM algorithm for parameter estimation (which is not sensitive to starting values), optimisation of matches by means of a linear sum assignment program, and a probability model that addresses both m and u probabilities for all value states of a field (Jaro, 1995).

3.2.3 A practical approach

The standard approach to starting a new linkage job is to initially make use of frequency ratios obtained in a previous job. After running the matching algorithm, a clerical check of borderline links is carried out, and then numerators (based on the file of linked pairs), and denominators (based on the file of randomly matched unlinkable pairs) are updated.

Updating for each new job would not be necessary if we could assume that the quality of the data in the file initiating the search was essentially the same as that in the file being searched. However, this is rarely the case, so a file of randomly matched unlinkable pairs should be created routinely as part of each linkage job.

Full comparison of every record in the file initiating the search against every record in the file being searched would render linkage impractical for all but very small files. To avoid such unnecessary work, crude methods of comparison are used first, to exclude the majority of candidate pairs for linkage.

A common strategy is:

(i) Block file on phonetic code of surname and date of birth (to avoid creating most of the potential pairs)

(ii) Examine key identifiers, for example initials or date of birth and exclude pairs where there is no agreement

(iii) Examine all available identifiers for remaining pairs

(iv) Use clerical checking for pairs with borderline overall odds in favour of a link.

Blocking involves partitioning both files into mutually exclusive and exhaustive subsets, and only looking for matches within each subset. To avoid missing links because of errors in a blocking variable, it is necessary to make two blocking searches [for example, block on Soundex and block separately on date of birth]. Comparison of all identifiers can be cut-off when a record pair is seen to be unlinkable. A fixed threshold is implied for the complete accumulated frequency ratio. Less favourable accumulated ratios are tolerated until the remaining steps cease to hold any prospect of restoring the balance to an acceptable level at the end. For comparisons with a small number of outcomes, it is logically simpler to make the comparisons value-specific from the beginning rather than carrying out a conversion step later.

When there may be several records relating to the same person, problems of competing linkages may be lessened by carrying out internal linkages within the two files prior to the external linkage of the two files, and then linking the whole patient group to the external search record.

There are 6 main sources of linkage error:

1) Disagreements on the blocking identifiers, for genuinely linkable pairs.

2) Inappropriately placed thresholds resulting in unnecessarily high numbers of errors.

3) Lack of discriminating power in the identifiers common to a record pair.

4) Underuse of the discriminating power - too few outcome levels.

5) Correlation between identifiers.

6) Imbalance between false positives and failures to achieve potential good links.

It has been suggested that where linkage results are to be used only for statistical purposes, false positive and false negative links are tolerable so long as they balance out one another. Thus, it may be adequate to simply place the threshold in the middle of the doubtful links. The problem arises in actually implementing this suggestion, since the only way to ensure that the false positives and false negatives are balanced is to verify the borderline links. While a balance may be sufficient when looking at counts of events, it is not appropriate when outcome records are to be related to external variables. An alternative approach is to carry out statistical analyses on results using three different thresholds. The middle one will be the likely best threshold with the other two being deliberately set too high and too low. If the same statistical associations hold for all three thresholds, then they are unlikely to be due to an imbalance between the false positives and false negatives.

3.3 Alternative models

The Tepping model for record linkage (Tepping, 1968) uses the same underlying framework as the Fellegi-Sunter model, but has no restriction on the number of possible decisions and makes 'cost' an explicit element of the model. 'Cost' here relates to both the costs incurred by decision D_2 leading to more clerical checking being required, and the cost of losses associated with matching errors. For each decision D_i, cost is assumed to be a function of the conditional probability of a match, and the linkage rule is: For any $P(match|\gamma)$, choose the D_i with the smallest cost. The main consideration in choosing between the Tepping and Fellegi-Sunter models would be the feasibility of estimating

the parameters, which are different for the two models (Jabine and Scheuren, 1986). The idea of a model involving minimisation of a cost function was first introduced by Nathan in 1967, but the simple model he proposed assumed that the information used for matching was complete and invariant, which is rarely the case.

Some other alternatives to the subjective probability model described by Fellegi and Sunter in 1969 were discussed by Copas and Hilton in their 1990 paper. Their work was based around the linkage of information on non-EEC citizens arriving in the UK with information on those leaving the UK, as part of a project for the Home Office.

The hit-miss model (Copas and Hilton, 1990) envisages a binary trial which results in a hit (where identifying information agrees exactly) with probability 1-a and a miss with probability a. For a given true record T, a hit results in X=T, but for a miss, the observed record X is randomly distributed over all possible values in proportions similar to the overall incidence of true values (a miss may by chance lead to the record being correct).

$$P(X=i|T=j)= \alpha_{ij}=\begin{cases} a\beta_i & i \neq j \\ 1-b-a(1-\beta_i) & i = j \\ b & i \text{ or } j = \text{blank} \end{cases}$$

where $\beta_i = \Sigma_j p_{ij}$, estimated by the overall relative frequency of value i among non-blank records, and b=probability that T is recorded as blank.

The model asserts that the off-diagonal cells in the n×n contingency table for matched pairs show a pattern of independence.

The log-likelihood ratio for the pair (i,j) under the hit-miss model is

$$\begin{cases} \log c + 2\log(1-b) & i \neq j \\ \log\left\{1-c(1-b)^{-2}(1-\beta_i)\right\} - \log \beta_i & i = j \\ 0 & i \text{ or } j = \text{blank} \end{cases}$$

where c=a(2-a-2b). Therefore, a pair with one or both records missing gives no matching information, and all discordant pairs give the same likelihood ratio. In practice, double blanks should often be thought of as 'agreements' since a person may genuinely not know an item of information. This can be modelled by taking b as a random effect, which results in double blanks giving evidence for a match, and single blanks giving evidence against a match. The levels of agreement approach, as described earlier, which is most closely

related to the hit-miss model is simply to note whether the pair of records agree or disagree. The use of adjustment factors to update global frequency ratios to value-specific frequency ratios (see section 3.2.1 and Newcombe, 1988) gives likelihood ratios which are similar, although not identical, to the ratios achieved by fitting a hit-miss model.

Use of extended models allows the strong assumptions of symmetry to be relaxed, but, if they are to be useful, they should only involve a relatively small number of extra parameters. Further examples of more specific models are discussed in Copas and Hilton (1990), including models which address the problems of highly correlated identifiers.

An alternative to a subjective probability model would be to use an expert system, although the added complication is unlikely to improve the accuracy of record linkage to any significant degree, in practice. Research into the use of neural networks in this area is currently being carried out by Gill and colleagues at the Oxford Record Linkage Study.

3.4 Medical record linkage in Scotland

For many years, Scotland has held sets of computerised medical records at a national level. In 1968, Heasman first outlined the potential for bringing these records together into patient groups using sex, surname, initials and date of birth (Heasman, 1968). He proposed linking intra-hospital and death records first, but thought it might be possible to incorporate cancer registrations and mental hospital records at a later date. The problems he anticipated at this stage included variation in identifying details on the records relating to the same person (leading to possibly 1 in 20 matches, or failures to match being incorrect) and issues of medical confidentiality. The potential for mismatches meant that this system could not be used to access records for use in the clinical care of a patient or for administrative decisions concerning individuals.

Throughout the next decade a system was set up to collect together medical records for an individual, and to enable the linked data to be used for epidemiological and health services research (Heasman and Clarke, 1979). The Scottish Record Linkage System (SRL), which

is based at the Information and Statistics Division (ISD) of the National Health Service in Scotland, uses only records which are required for routine health statistical purposes. There is no extra data collection involved. The patient matching depends on the fact that less than one per cent of individuals have the same surname, initials, sex and date of birth, out of the Scottish population of approximately five million. At this stage, a new linked file was specially created for each project, making the procedure expensive and time consuming.

In 1989, development began of a new system in which all records for individual patients would be permanently held together in a linked data set (Kendrick and Clarke, 1993). At present, the morbidity data set holds hospital discharge (SMR1) records since 1981 together with Scottish Cancer Registry records (SMR6) and Registrar General's death records - over 12 million records in total. Another linked data set holds patient groups of psychiatric hospital admission (SMR4) records since 1970. Other records are not on a linked database, but are available for ad hoc linkage to other files.

3.4.1 Scottish national data sets

The national data sets which will be used in this thesis are now described.

SMR6

The Scottish Cancer Registration Scheme had its origins in a system set up in 1936 by the National Radium Commission to record the progress of radiotherapy patients. It is now one of a number of Scottish Morbidity Record (SMR) Schemes, and information is collected on a standard form known as an SMR6 (see Appendix C). The scheme registers individual primary tumours, which must be registered separately even when they occur on the same individual (Common Services Agency, 1990). Scotland is divided into 5 regional registries, which submit data to the Information and Statistics Division (ISD) of the Common Services Agency of the Scottish Health Service, in Edinburgh. Notification comes to the registries from pathology departments, haematology departments,

radiotherapy departments, hospital medical records departments, the General Register Office (post-mortem information) and nursing homes.

An attempt has been made to validate central registration data on all childhood leukaemia cases in Scotland between 1968 and 1981 (Glass et al, 1987). It was found that 44% of these registrations contained minor errors, including 41 minor diagnosis differences out of the 629 records on the cancer registry. A further 6 records had wrong diagnoses and 8 leukaemia cases identified from hospital notes did not appear on the cancer registry.

SMR1

Scottish Hospital Inpatient statistics come from SMR1 forms (see Appendix C) which have been completed since 1961 for all inpatient and day case discharges, deaths or transfers from Scottish non-psychiatric, non-obstetric wards. Since 1989, information from the SMR1s has been used to plan the financial management of hospitals, but their completion is often delegated to clerical staff and unsupervised by clinicians.

An audit of the accuracy of recorded SMR1 data for a single ENT operation (Denholm et al, 1993), taking the theatre book completed twice daily by the sister-in-charge to be the gold standard, found that the SMR1 had the correct operation code for 88% of these operations performed, although the coding errors mainly related to other ENT procedures and were not too serious. A review of a sample of discharges in Greater Glasgow Health Board in 1987 for gastrointestinal diagnoses (Kohli and Knill-Jones, 1992) identified minor disease coding errors for 3.8% of cases, intermediate errors for 0.8% and major errors for 0.8%.

SMR4

Scottish Psychiatric Inpatient statistics come from SMR4 forms (see Appendix C). The current version of the form came into effect in 1986. These forms are in two parts. The first part is submitted to ISD on a patient's admission, and the second part is completed on discharge.

SMR20

In 1981, the SMR20 (see Appendix C) scheme was set up to provide accurate information on the position of cardiac surgery in Scotland. The Scottish Cardiac Surgery Register collects information on patients who are on the waiting list for cardiac surgery, as well as those actually admitted to hospital, and enables quarterly checks to be carried out on the extent of cardiac surgery waiting lists. Consultants in Public Health Medicine in each Health Board area are responsible for the local collection of data for the Register (Common Services Agency, 1990).

Other records

ISD also holds a file of death records which can be used in linkages. This data does not 'belong' to ISD however, but to the Registrar General for Scotland. ISD receives an update file from the General Register Office (GRO) at 3-monthly intervals.

3.4.2 Methods in use in the SRL system

The basis of forming a linked data set is the comparison of 2 records, and the decision as to whether or not they relate to the same individual. For each of the main items of identifying information used to link records there may be a discrepancy rate of up to three per cent in pairs of records belonging to the same person due to errors in recording, so exact matching could miss many true links. To allow for imperfections in the data, the Scottish Record Linkage System uses methods of probability matching (Kendrick and Clarke, 1993) which have been developed over the past 30 years in Canada (Newcombe, 1988), Oxford (Acheson, 1987) and in Scotland itself (Heasman and Clarke, 1979). Probability matching is useful as all computer records are liable to error or variability - people move house, and women change their surname when they get married, in addition to differences in spelling of names and errors in writing down data, deciphering hand-writing and at data-entry. In the medical context, errors may occur through misrecording of information due to conditions of stress. These errors mean that two records for the same person may not agree, while two records which do agree may relate to different people.

As well as maintaining the linked datasets described earlier, the Scottish Record Linkage System provides facilities for ad hoc linkages with other data sets available at ISD, or the linkage of external files to the linked datasets. An external dataset could consist of, for example, identifying information for patients being followed up as part of a clinical trial. New data are linked into the main linked databases on an annual basis, and for this reason the databases could not be used as an administrative system for tracking individual patients, as well as the probabilistic basis of the linkage techniques leading to a small proportion of mismatches. Records are stored together in a flat file, in chronological order, retaining their original unlinked format preceded by fields of linkage information such as a unique personal identifier for each patient group, and a marker indicating which SMR1 records refer to the same period of continuous in-patient stay. This stay marker is useful since separate SMR1s are completed for transfers between hospitals, specialties and consultants. The holding of records in patient groups is also useful for carrying out subject-based rather than episode-based analysis. Since the different types of records have different formats, they are usually accessed via FORTRAN programs.

The computer matching algorithm calculates a score for each pair of records, which is equivalent to the odds that they belong to the same person. The overall score is the sum of the scores derived from the comparison of each item of identifying information. The comparison scores are weighted according to the rarity of the information. For example, if both records have the first initial J, the score is increased by a small amount, but if both records have the first initial Z, the score is increased by a larger amount, reflecting the fact that agreement on an uncommon initial increases the probability of a match much more than agreement on a common initial. Similar negative weightings are applied to the level of disagreement between items. Items of identifying information contributing to the score should be statistically independent as far as possible.

The basic core of identifying information used is:

1) Surname (and it's phonetic code to overcome differences in spelling)
2) First initial (also full forename and second initial when available)
3) Sex
4) Year, month and day of birth
5) Postcode

The phonetic code used at SRL for the surname is the Soundex code. It has been found to remain unchanged in about $\frac{2}{3}$ of the spelling variations observed in linked pairs of records (Newcombe, 1967), and it sets aside only a small part of the total discriminating power of the full alphabetic surname, although it is less satisfactory for Oriental names. The Soundex coding system reduces a name to a code consisting of the leading letter followed by 3 digits. All vowels after the first letter are ignored, as are W and H. The remaining letters are coded as 1 (B,P,F,V), 3 (D,T), 4 (L), 5 (M,N), 6 (R) and 2 (all other consonants). If the name has fewer than 3 coded letters, trailing zeros are added. A Soundex weight is assigned to each code, reflecting the rarity of the Soundex code throughout the Scottish population as a whole. Soundex codes applying to relatively common surnames will have low weightings, whilst codes for rare surnames will have the maximum weighting of 15.00. The Soundex weight is used by the computer matching algorithm in the calculation of the comparison score.

Additional identifying information could be used for specific types of linkage, for example, hospital case reference number could be used for internal SMR1 linkages, or date of discharge could be compared with date of death for linkage of SMR1 and death records.

In order to cut down on the number of comparisons required, the records are blocked on
1) Phonetic code of surname and first initial
2) Date of birth
Full comparison is only carried out for records which agree on either of the blocking criteria.

A common problem is that people sometimes use their middle name for everyday purposes, and this is entered as forename on hospital records, so cross-comparison of first and second forename initials is also carried out.

Each data set has its own problems, so the distribution of scores varies for different linkages. The threshold (that is, the score at which the decision to link is made) is determined by clerical checking of a sample of pairs of records, for each linkage carried

out, and is usually placed where the implied odds in favour of a correct match are 50:50 or better.

The system is designed to operate on the basis of the computer matching algorithm as far as possible, with minimal clerical checking. It will be described further in the next chapter, in the context of a practical linkage exercise involving the West of Scotland Coronary Prevention Study. Specific groups of records are, however, targeted for clerical monitoring, for example subjects with 2 death records, or with a hospitalisation occurring after a death record. Clerical checking of the main linked database has shown that both the false positive (incorrect links) and false negative (links missed) rates are around 1 per cent for internal ISD linkages where the national data sets are linked into patient groups. Since each record in a patient group will have been compared with every other record in that patient group, the likelihood that one or more records in the group does not correctly belong to that group increases as the group size increases due to the cumulative errors of each pairwise comparison. Clerical checking would improve the quality of the linkage even further, but would require much time and expense. Computerised record linkage is valuable if it identifies a high proportion of true events of interest, whilst keeping clerical checking to exclude 'false positives' to manageable proportions. Also, it's cost should not exceed that of alternative methods of follow-up.

3.5 Record linkage outside Scotland

Canada

The earliest probability linkages were carried out as an experiment by Atomic Energy Canada in the 1950s, in an attempt to keep track of individuals who had been exposed to low-level radiation (Newcombe et al, 1959). S.J. Axford of Statistics Canada first suggested phonetic coding of surnames, and blocking records by surname. A decade later, the linkage rationale was restated and expanded into a formal mathematical theory by Fellegi and Sunter at Statistics Canada (Fellegi and Sunter, 1969) as described earlier. Close scrutiny of actual linked and unlinkable files provided insights which would have been missed if refinements had only been sought through developing theory. Statistics

Canada has now developed an all-embracing system to deal with many diverse linkage jobs involving the Canadian mortality database which extends back to 1950. This system was previously known as GIRLS (Generalised Iterative Record Linkage System) and described in detail elsewhere (Howe and Lindsay, 1981). It is now known as CANLINK. This system involves iterative updates of outcome frequencies for new linkages and the setting of two subjective thresholds to separate links, non-links and possible links.

California

In 1981, the California Automated Mortality Linkage System (CAMLIS) was set up to facilitate the conduct of follow-up studies in California (Arellano et al, 1984). It uses probabilistic linkage decision criteria to perform mortality searches. CAMLIS offers a rapid, accurate and cost-effective means of accessing the California state mortality files and has been shown to have both reasonably high sensitivity (0.89-0.97) and specificity (0.93-0.99).

The Netherlands

Probability linkage has also been carried out in the Netherlands, although a formal system has not been set up for running linkages repeatedly. For example, an assessment was carried out to compare the completeness of cancer registration on the Limburg regional cancer registry and the GP's centralised database, using record linkage techniques (Schouten et al, 1993). If the information on the 2 databases differed, it was verified using the source forms at the registry and by contacting the GP concerned. By combining information from the 2 files, it was determined which malignancies should have been registered. The cancer registry had recorded 307 out of 319 eligible malignancies (96.2%), with 5 of the 12 missed being due to systematic shortcomings in the notification procedures.

Oxford

The Oxford Record Linkage Study (ORLS) was founded by E.D. Acheson in 1962, as a pilot study to investigate the feasibility and cost of collecting records about medical events and linking these records, by computer, on a person and family basis, develop computer methods of record linkage, study applications of the files in medical and operational research, and if successful, promote it's extension on a national basis

(Acheson, 1987). Data collection began in 1963 and covered around 350,000 people. By 1982, this had been extended to the whole of the Oxford region and part of Berkshire, (approximately 2.3 million people). Data are collected on general, maternity and psychiatric hospitalisations, births and deaths, and linked together using methods similar to those employed by the Scottish Record Linkage System, as described earlier. The linked database is particularly useful for analysis which requires unduplicated counts of people and studies of successive events since all the records are grouped into patient record sets (Goldacre, 1986), but extension to a national database has proved impractical in England and Wales.

Chapter 4

Record Linkage and WOSCOPS

4.1 Record linkage as a method of subject follow-up

The existence of Scottish national databases covering all deaths, incident cancers, hospitalisations, psychiatric hospitalisations and cardiac surgical procedures provides the opportunity to use record linkage techniques as an alternative method of follow-up for patients who have been randomised into clinical trials or are members of a screened cohort.

4.1.1 Clinical Trials

In a hospital based clinical trial the patients are well motivated and it is relatively easy to maintain contact with them and obtain information on adverse events occurring during the trial. Studies conducted in the community, especially primary prevention trials, often have very large numbers of subjects selected from a wide population base, making it more difficult for investigators to maintain contact with each subject. In primary prevention trials, the subjects are generally disease free and hence are less well motivated to comply with the study protocol. In many cases, follow-up is carried out independently of the subject's general practitioner and thus the quality of the data collected substantially depends on the goodwill of the patient in attending visits and reporting events. Poor attendance at follow-up visits can lead to incomplete reporting of

adverse events and study endpoints. In a clinical trial, withdrawals and defaulters will not necessarily be spread evenly among the study treatment arms, leading to the possibility of bias in the reporting of efficacy and in the adverse event profile of the treatments being tested in the study. Withdrawal rates can be high. For example, by the end of the Helsinki Heart Study (Frick et al, 1987), approximately 30% of the subjects had withdrawn from study medication. Although many of these subjects may have maintained regular contact with the study, there is clearly greater potential for incomplete reporting of clinical events in this group of subjects. In some countries (including Scotland) there are national death and cancer registries with which subjects can be flagged to reduce the effect of the problem of under-reporting. However, non-fatal events, which may be of particular interest in some studies, are not always available. The cost of maintaining completeness of follow-up and documentation of serious adverse events, particularly hospitalisations, is also a major limiting factor in the design of large-scale clinical trials. Any method which can reduce costs and yet retain adequate levels of reporting is worthy of serious consideration.

Record linkage provides an alternative, cost-effective method of subject follow-up using national databases, which is unaffected by the subject's compliance with the study protocol, and attendance at trial visits. It does not require any further contact with the patient once identifying information has been recorded.

4.1.2 Epidemiology

As is the case for clinical trials, in epidemiological studies the cost of patient follow-up can be prohibitive, although a baseline visit for the measurement of risk factors may be feasible. Computerised record linkage systems can be used to search databases of outcome data collected for another purpose (for example, routine Health Service data in the case of the Scottish Record Linkage System) for the study participants, as an alternative to active follow-up of the individual subjects. However, this is only appropriate if such systems can be validated. Computerised record linkage is particularly

cost-effective when a large number of subjects is involved, as increased number of subjects makes only a small difference to the costs involved in record linkage, whereas the costs of individual subject follow-up would increase substantially. Neutel et al (1991) recommend the use of record linkage procedures to reduce the cost and time required by follow-up studies. Cost obviously increases with the number of subjects involved, the duration of follow-up and the size of the area covered.

The advantages and disadvantages of using record linkage techniques in the epidemiological context are similar to those in the context of clinical trial follow-up.

The advantages include:

1) Cost-effectiveness - epidemiological studies frequently require data on many subjects, collected at different points in time. While record linkage is not cheap, it is considerably cheaper than the alternative methods of follow-up.

2) Fewer subjects lost to follow-up - record linkage will continue to identify records relating to subjects who have moved house (although the certainty of subject identification will be lower), or simply lost interest in the study.

3) Reduction in recall bias - since the outcome data is produced routinely by hospitals and does not rely on the subject's memory.

4) Unobtrusive data collection - since direct contact with the subject is no longer required in order to obtain accurate information on outcomes.

5) Opportunities to consider whole populations and to obtain results quickly for any new hypothesis of interest, since a database covering many outcomes is readily available and there is no need to wait for more data to be collected.

The disadvantages include:

1) Inflexible data - since the data being linked to was originally collected for another purpose, it may be impossible to get information on the variables of interest.

2) Subjects lost to follow-up - missed linkages due to errors in identifying information, subjective positioning of the threshold linkage score or events not

detected on the database being linked to (for example, events occurring outside the area covered by the database). This will be important if the subjects lost to follow-up differ in exposure and outcome from the rest of the subjects involved in the study.

3) Data overload - the ease of data acquisition and analysis can lead to 'trawls' of the data in a search for significant results.

4.2 Aims of this project

The primary objective of this project was to set up computerised record linkage systems for the WOSCOPS subjects.

The first system was to be used for the 6595 randomised subjects and encompass both routine linkages to the linked databases held at the Scottish Record Linkage System (SRL), and ad hoc linkages to unlinked databases of more recent records.

The second computerised record linkage system was to be used for the cohort of subjects who attended an initial screening visit for the WOSCOPS study. Although around 105,000 subjects were screened, only 97,165 had sufficient subject identifying information to make follow-up by record linkage possible.

These record linkage systems were to be set up so that it would be relatively straightforward to re-run the linkages repeatedly as more years of follow-up were accumulated. This was considered an important aspect since the more meaningful epidemiological data involves many years of follow-up and it's acquisition will extend beyond the end of this particular project. Cancers, in particular, may take many years to develop.

For the clinical trial arm of the project, a comparative study was to be carried out, in which adverse events obtained by record linkage for the 6595 randomised subjects would be compared with the events on the WOSCOPS database of adverse events identified by individual subject follow-up. The aim was to provide an assessment of the

effectiveness of each of these methods of subject follow-up. Further ad hoc linkages were to be carried out to 1994 and 1995 records where available, to provide additional information on events, for the WOSCOPS randomised subjects. However, only data up to the end of 1993 was to be included in the comparative study, in order to guarantee adequate time for the data to have been reported by the subject and to pass through the WOSCOPS reporting system.

For the epidemiological branch of the project, computerised record linkage provides the only feasible method by which adverse event information can be obtained for the WOSCOPS screened cohort. The adverse events identified by record linkage for the screened cohort were to be analysed in order to investigate the relationships between baseline risk factors and coronary heart disease, cancer, trauma and all-cause mortality as described in chapter 1. Analysis was to focus mainly on the group of 45-64 year old men who were targeted by the screening process. There was also interest in the screened women on their own since the majority of previous studies of coronary heart disease have focused on middle-aged men and it is unclear whether the same relationships among baseline risk factors and patterns of disease are applicable to women. Laboratory data such as fibrinogen levels are currently of great interest in relation to coronary heart disease and were to be examined for associations with mortality outcome determined via record linkage. Laboratory measurements are only available for the approximately 13,600 men who reached screening visit 3. The main constraint on the analysis possible for the screened cohort was the need to maintain subject anonymity. Adverse event data were available only by a category of baseline risk factors and was not identifiable at an individual subject level.

The final objective of this project was to compare the event rates observed in the screened cohort with the event rates observed in the general population in the area of screening as defined by postcode sectors.

4.3 Author's contribution to this project

The specific contribution made by the author to addressing the aims of this project were as follows:

1. Co-ordination of the postcoding exercise for the 6595 randomised subjects and the 97,155 addresses available for the screened cohort. This involved address editing prior to sending a file of addresses to Greater Glasgow Health Board for postcoding, liaison with a postcoding bureau for the 29,310 screenee records not coded by the Health Board, and supervision of the manual postcoding of the records not coded by the Health Board or bureau (see section 4.4). All known addresses for each randomised subject were postcoded in order to maximise the identifying information which would later be available for record linkage.

2. Production of files of linking information for the randomised subjects and the screened cohort. A duplicate record was produced for any change in recorded identifying information for a randomised subject (see section 4.4).

3. Development of one-pass linkage programs (see section 4.5)
a) to take advantage of characteristics of the WOSCOPS data sets and
b) for use with other SMR records [one-pass programs had previously been used only for SMR1 and death records].

4. Linkage of randomised subjects to linked databases of SMR1, SMR4, SMR6 and Registrar General Death records until the end of 1991. These linkages were repeated for SMR1, SMR4, SMR6 and Death records until the end of 1992, and then for SMR1, SMR4 and Death records until the end of 1993, as these years of data became available. The randomised subjects were also linked to SMR20 records until April 1994. This was achieved on the first linkage run since the SMR20 database was on-line (that is, updated to current date). The SMR databases were described in section 3.4.1.

5. Ad hoc linkages of the randomised subjects to interim, unvalidated files to provide warning of events occurring in 1994 and 1995. These ad hoc linkages were carried out to obtain the most up-to-date information possible towards the end of the follow-up period for the clinical trial branch of WOSCOPS. Setting up this system involved modification of programs used by staff at SRL, for use with unlinked databases. However, a full comparison to assess the completeness of each method of follow-up was not carried out for these later ad hoc linkages, since the completeness of this data could not be guaranteed. They are thus not discussed in any detail in this thesis.

6. Linkage of the screened cohort to the linked database of SMR1s until the end of 1993, deaths until the end of 1993, and SMR6s until the end of 1992. This exercise involved further adaptation of the programs used at SRL, to accommodate the large number of subject records since there was insufficient computer memory to store all the WOSCOPS records at once and carry out each linkage in a single run.

7. Comparative study for the randomised subjects. The comparison of the adverse events identified by record linkage and by individual subject follow-up allowed an assessment of the advantages and disadvantages of these two methods of subject follow-up in the context of a clinical trial, and an assessment of the accuracy of the probabilistic record linkage process employed by SRL. The comparison exercise is in some ways similar to carrying out a second linkage exercise, linking specific events on 2 separate databases (WOSCOPS and SRL identified events) as opposed to linking subjects to their own records on the Scottish national databases as described in Chapter 3. This comparative study was carried out twice. Records until the end of 1991 were compared in the first instance (West of Scotland Coronary Prevention Study Group (in press)). The comparison exercise was then repeated for cancer registrations until the end of 1992, deaths and hospitalisations until the end of 1993 and cardiac surgery registrations until April 1994 (see chapter 5).

8. Development of procedures for the investigation of events identified by linkage but not currently on the WOSCOPS database (in consultation with a clinical research

associate), and keeping track of these events throughout the investigation process. These events were investigated to verify whether or not they did truly belong to the WOSCOPS subject, and , if so, whether there was any reason why the event was missed by WOSCOPS. It was hoped that previously unidentified adverse events, would mainly relate to subjects who had withdrawn from taking trial medication or who had been defaulting from their scheduled trial visits. Procedures were then developed for the documentation of new events found by linkage (in consultation with a clinical research associate and staff at the Data Centre). The database of adverse events identified via record linkage was thus used both to support and to supplement existing WOSCOPS adverse event reporting, so that the national databases enhanced the validity of the final serious adverse event report on the WOSCOPS clinical trial.

9. A comparison of the linkage quality achieved with screening visit 1 data and with updated prepped data (see section 5.4.1), and a comparison of the linkage quality achieved with postcoded and unpostcoded data, for the randomised subjects for deaths until the end of 1993 (see section 5.4.2).

10. Management of resolution tables for all linkage derived records brought to the Data Centre for the randomised subjects.

11. Production of categorisations by which the events identified by linkage for the screened cohort could be extracted and analysed whilst maintaining subject anonymity (see chapter 6).

12. Analysis of data for the screened cohort (see chapter 6). This analysis by category was carried out for mortality in three groups of subjects:

- approximately 80,000 45-64 year old screened men, by a general category of CHD risk factors
- approximately 14,000 screened women, by a general category of CHD risk factors
- approximately 13,000 men who reached screening visit 3, by a category based on laboratory measurements.

Analysis was also carried out for the hospitalisations and cancer registrations for the group of 80,000 screened men.

13. Examination of mortality rates observed in the population in the area of screening (as defined by postcode sectors) and calculated from data on deaths in the general population, obtained from the Registrar General for Scotland, along with the 1991 census data. These mortality rates in the screening area population were compared to the rates observed for the screened cohort in order to address the question of how representative a screened cohort is of the general population, in terms of mortality rates, (see chapter 7).

4.4 Preparation of WOSCOPS data for linkage

In order to carry out record linkage, it was necessary to first produce files of identifying information for the 6595 randomised subjects and the 97,165 screened subjects. The identifying information used in the WOSCOPS linkages was

1) Surname (and it's phonetic code - Soundex)
2) First initial (or full forename and second name depending on the file being linked to)
3) Sex
4) Day, month and year of birth
5) Postcode of home address

For the 6595 randomised subjects, these data were extracted from the computer file containing data entered from the WOSCOPS SC1 form (see Appendix A). The file of linking information contained a duplicate record for each known address for a subject, enabling more accurate linkage to SRL records occurring at different points in time. Prior to producing this file, a postcoding exercise was carried out, as the postcode had

not been obtained consistently. For the 6595 randomised subjects, every known address for the subject was postcoded (that is, baseline address, and any notified changes of address). These records were sent to Greater Glasgow Health Board where an auto-postcoding algorithm succeeded in postcoding approximately two thirds of the addresses. The remainder were postcoded manually by WOSCOPS Data Centre staff.

For the cohort of 97,165 subjects who attended an initial screening visit the file of identifying details was based on the information recorded at this visit. As for the randomised subjects, a postcoding exercise was carried out for the 97,155 screenee records for which an address was available. 10 subjects had no address but were included in the linkage since name and date of birth information was available. On this occasion, Greater Glasgow Health Board succeeded in autopostcoding 67,845 records (approximately two thirds as before). The remaining 29,310 records were sent to Chester for bureau postcoding. The bureau successfully postcoded 27372 addresses, and provided a breakdown of the completeness of their postcoding (see Table 4.1). As a quality control check, the first 100 addresses which had been sent to the bureau were clerically checked by a member of the Data Centre staff. It was found that 91 addresses had been postcoded correctly, 1 had not been postcoded and 8 had been assigned a doubtful postcode. The 1937 UK addresses still remaining without a full postcode were postcoded manually by WOSCOPS Data Centre staff.

Postcode Status	Number of records
Full postcode	27372
Partial postcode	951
No postcode - UK	986
No postcode - foreign country	1
Total	**29310**

Table 4.1 Completeness of the postcoding carried out by the bureau. This was for the 29,310 addresses relating to the screened cohort which had not been successfully postcoded by Greater Glasgow Health Board.

4.5 Record linkage at SRL for WOSCOPS subjects

The computerised record linkage exercise was carried out using linkage programs written by staff at SRL and designed for the linkage of external data to Information and Statistics Division (ISD) datasets. These programs were modified to take advantage of the characteristics of the WOSCOPS datasets. A general account of the principles involved is given elsewhere (Scottish Record Linkage Team, 1995). Linkage was based on the one-pass method, in which the WOSCOPS records were read into memory and blocking was carried out using indexes, so that the records on the linked database could each be considered sequentially, and this huge dataset didn't need to be re-sorted for each of the blocking criteria used (Scottish Record Linkage Team, 1995). Each record from the linked database was compared to each WOSCOPS record which agreed on the blocking criteria (for example, phonetic code of surname) and a linkage score was accumulated by adding and subtracting various adjustment weights depending on the level of agreement or disagreement on the items of identifying information (see chapter 3 for further details on the record linkage process). The file of linked pairs from which these weights were derived was produced at ISD using progressive refinement of the linkage algorithm. These weights may be amended according to features of the particular datasets being linked. For example, for the WOSCOPS randomised subjects, a large deduction was made for differences in the phonetic code of the surname since the dataset contained men only. It is unlikely that surname would change for men, while it usually changes for women on marriage. Thus, agreement on surname carries more discriminating power for men than it does for women. In practice, this 'fine-tuning' is often done fairly subjectively. A patient record set from the linked database cannot be linked with more than one WOSCOPS subject - known at SRL as the 'best link' principle. However, conversely, more than one patient record set can be linked with a given WOSCOPS subject, initially. These groups are then split, taking the patient record set with which the highest score is achieved and retaining links only when this score achieves the revised threshold score set by clerical

checking. The best link principle is thus applied in both directions.

The main stages of the record linkage process used were:

1. Standardisation of data

– involving assignment of Soundex codes and weights to surnames, and formatting of the postcode on both the WOSCOPS and the linked database records.

2. Sorting

– of the WOSCOPS records into blocks as appropriate, for example, Soundex code order.

3. Linkage algorithm

– the computer linkage algorithm was then run to perform the linkage. This algorithm is illustrated in Figure 4.1, and further details can be found in Appendix D.

4. Clerical review

– to decide on the appropriate score at which to set the threshold for separating links and non-links.

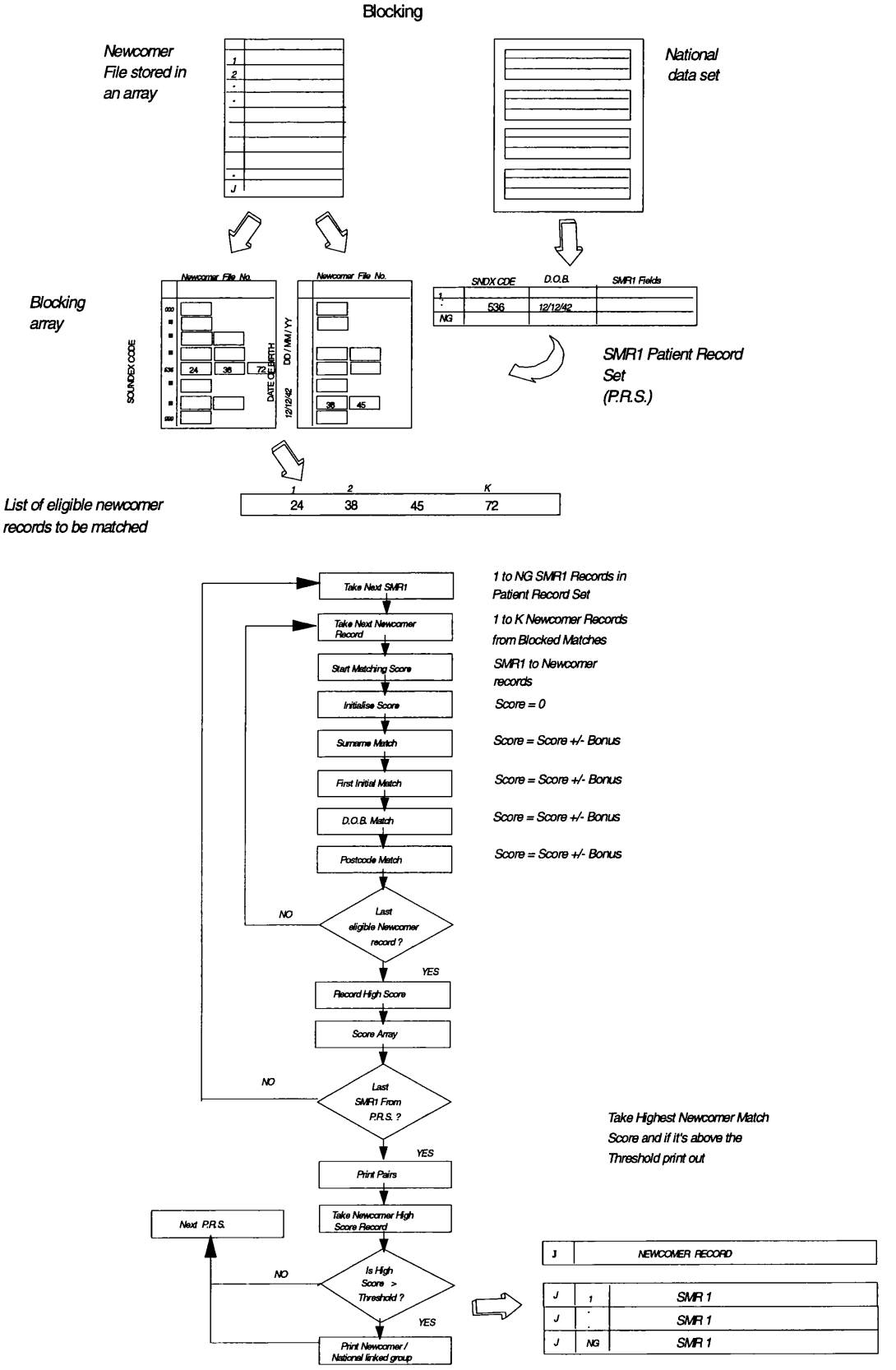


Figure 4.1 - Linking algorithm: flow diagram
(adapted from SRL team, 1995; and used by permission of SRL)

In general, the final score will be in the range -50 to 100, but the range may vary depending on the amount of identifying information available for use in the linkage process. For example, for the SMR6 linkage based on the items of identifying information available to WOSCOPS, the final score was in the range -39.4 to 57.29, with the threshold score set at 26 for the randomised patients linkage, while for the SMR1 linkage which uses only first initial rather than full forename, the final score was in the range -36.9 to 54.29, with the threshold score set at 25 for both the randomised and screened patients linkages.

Graphs are usually produced of the frequency of each score against the full score range (that is, -50 to 100) and a reduced score range (for example, 15 to 60) (See Figures 4.2 and 4.3).

On the graph for the full range of scores (Figure 4.2), the first peak relates to comparison pairs (that is, pairings of WOSCOPS record and ISD record) which agree on Soundex code but very little else. The second peak corresponds to record pairs which agree on date of birth (the other blocking variable) but on little else. The later, smaller peaks correspond to records which have some level of agreement on both Soundex and date of birth.

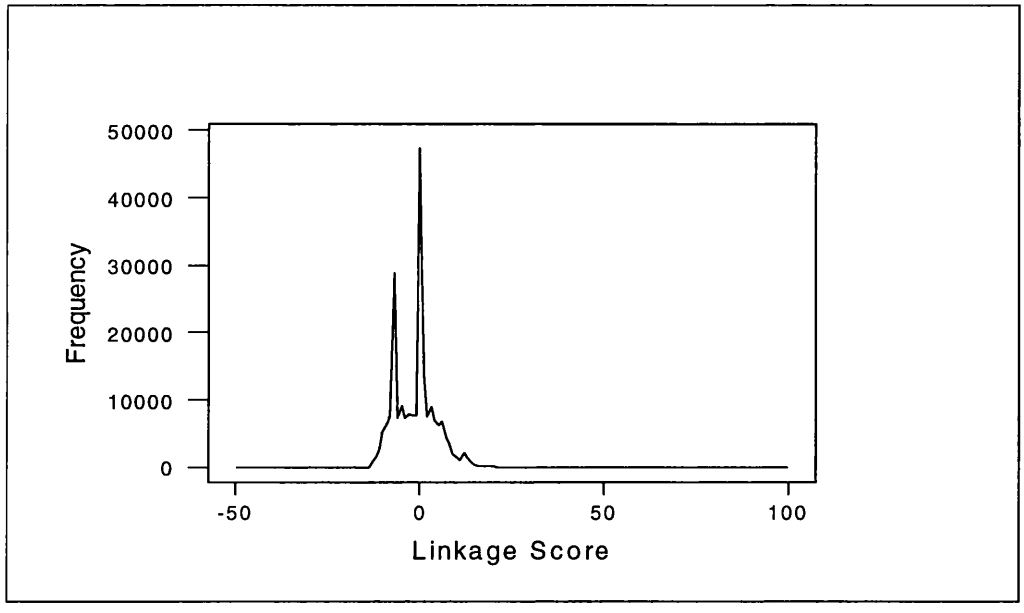


Figure 4.2 - Frequency graph of the score distribution for the full range of scores.

Since every blocked WOSCOPS-ISD pair of records are compared, many pairs of records are brought together which belong to totally different people - they will have low scores as they will have little identifying information in common, while pairs of records which have a high score will usually belong to the same person, since most of the identifying information must be identical. The huge number of compared record pairs which belong to different people means that in order to see the shape of the distribution for the much smaller number of links, a reduced range of scores must be considered.

On the graph for the upper range of the distribution (Figure 4.3), there is a clear trough between the 'good' and 'bad' matches. This graph is useful in deciding on a small range of scores (for example, 23 to 35) in which to carry out clerical checking to decide where the threshold score at which record pairs will be accepted as links should be fixed. After clerical checking the threshold was set at 30 for this linkage. The relatively level area between the threshold and the point at which the frequency starts to increase again represents pairs where the postcode did not agree, but a match was made on the basis of the other identifying information (name, sex and date of birth).

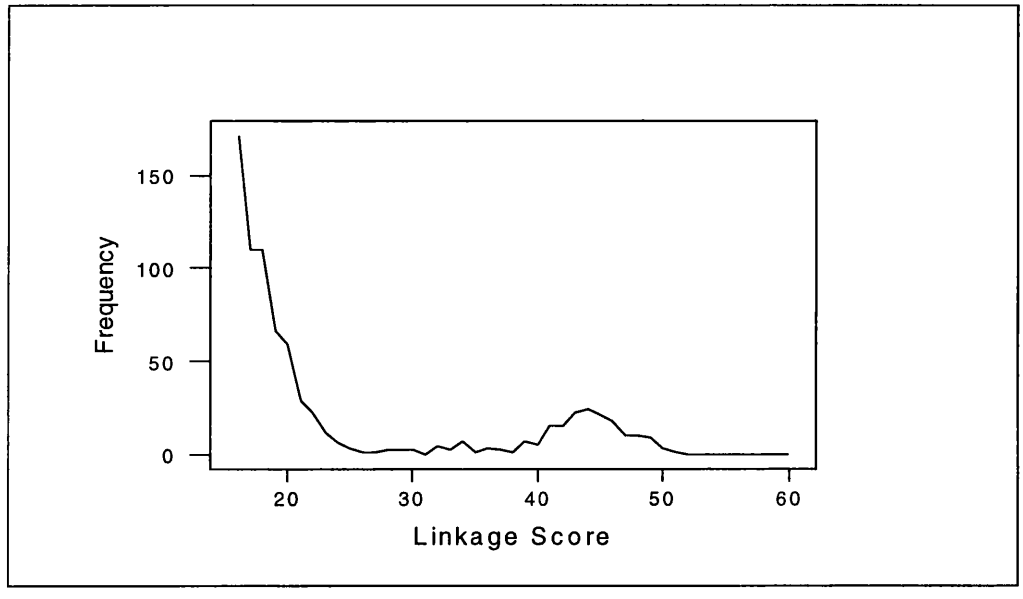


Figure 4.3 - Frequency graph of the upper range of the score distribution

The threshold point corresponds to the crossover point of two separate distributions, the distribution of scores for non-links, and the distribution of scores for links, making the overall score frequency distribution bimodal. A strongly bimodal frequency distribution with a deep trough at the optimum threshold level indicates that there is good discriminating power, with minimal overlap between the 'good' and 'bad' matches. So not only is there little uncertainty concerning the best position for the threshold, but the numbers of potential good links and actual bad links falling on the wrong sides of the threshold are small, when the identifying information has good discriminating power. While it is desirable that both errors be kept to a minimum, these are competing aims, and a conscious compromise is usually required.

4.6 Possible improvements

As will be seen in chapter 5, the linkage methods currently used at SRL are very effective and achieve a high standard of accuracy, despite the apparent simplicity of the methods employed.

An alternative to the single threshold approach employed in current ad hoc linkages at SRL would be to set two threshold levels, one on either side of a 'grey area', so as to separate 'good', 'doubtful' and 'bad' linkages as described by Newcombe, 1988. Clerical checking could then be employed to deal with records which fall in the 'doubtful' category. In the Oxford Record Linkage Study, the 'doubtful' category is further sub-divided into 'probably good' and 'probably bad' matches (Simmons, 1989).

SRL does not routinely update the frequencies contributing to the scores for new linkage exercises, as recommended by Newcombe in 1988. However, using scores derived through previous linkage exercises (with a little ad hoc 'tweaking') does not appear to have any great detrimental effect on the accuracy of results obtained.

Although surname is commonly used as a discriminator, forename is rarely utilised, with only the first initial, or exact agreement on the first four letters of the forename being considered. SRL does in fact have a rarely used look-up table of forename weights, and it would be possible to phonetically code (Soundex) forenames in a similar way to surnames. A further method would be to use a look-up table for nicknames, a possibility which has been explored in Canada (Newcombe, Fair, Lalonde, 1992).

Simmons, 1989 suggested using forename as a blocking variable, but experience from the linkage of WOSCOPS subjects to SMR6 records suggests that it may be worth relaxing the current block on Soundex code of surname+first initial to a block on Soundex code of surname alone. Where there are errors in date of birth (the other blocking variable) and the subject routinely uses a middle name, potential matches are not even considered due to the default blocking criteria employed at present.

Internal linkage errors (for example, when linking the SMR1 records into patient record sets) can lead to errors in external linkages (for example, SMR1s with WOSCOPS subjects) when whole patient groups of records are linked if one record pair achieves the threshold score. The pairings with other records in the SMR1 patient record set may in fact fall far below the threshold score. If clerical checking is to be kept to a minimum then it may be worth setting a second threshold value at which individual records within the patient record set would be excluded from linking to the external record, even though at least one record in the group achieves the threshold for group linkage to the external record. These internal linkage errors are routinely checked for at ISD for death records, where it is obvious that there should be at most one record for any subject. With record types where multiple records are possible it is more difficult to identify patient record sets likely to contain errors. If ad hoc external linkages are to be carried out before clerical checking of the linked databases is completed, then the external records should only be allowed to link with a single death record, as distinct from a single patient record set.

However, these possible improvements relate mainly to the specifics of the linkage

algorithm used, while the general method has been found to be straightforward and reliable, as will be shown in the next chapter.

Chapter 5

Record Linkage Comparative Study

5.1 Introduction

The aim of the record linkage comparative study is to compare computerised record linkage (CRL) with individual subject follow-up, in order to assess the advantages and disadvantages of these two methods of subject follow-up in the context of a clinical trial. The main outcome of this comparative study will be the production of tables containing a resolution for every adverse event record identified by either WOSCOPS independent follow-up or record linkage. Any records which can not be resolved as matching with a record on the other database will be investigated fully. The methods employed to carry out this comparative study are described below.

5.2 Methods

Events identified by computerised record linkage, carried out at the Scottish Record Linkage System (SRL), for the WOSCOPS randomised subjects, were compared with those recorded by the routine WOSCOPS system based on individual patient follow-up. Any event identified by computerised record linkage and not by the WOSCOPS system was followed up carefully to confirm that a WOSCOPS subject was involved and that a linkage mismatch had not occurred. This decision was based on direct contact with the

appropriate physician or surgeon in the hospital where the event had been reported and scrutiny of the relevant medical records by a clinical research associate. This resulted in a great deal of manual checking to assess the degree of completeness of each set of records. Similarly, events identified by WOSCOPS but not by record linkage were examined to see if there was any reasonable explanation for the failure to find a linkage match.

5.2.1 Comparison

As preparation of the files of linkage derived records, any records which occurred prior to the individual subject's randomisation date were removed. This was a necessary act since ethical permission for subject follow-up related only to the period after the subject had signed a consent form. It was also required for comparison purposes since WOSCOPS independent follow-up could only have documented events occurring after randomisation. SMR1 forms are completed for day case admissions as well as for admissions requiring overnight stays. However, day cases are not routinely documented by WOSCOPS, so any records with the date of admission equal to the date of discharge were excluded from both the WOSCOPS and linkage files. Care must be taken in the matching of events, particularly hospitalisations, since in the Scottish national system a new record is generated whenever a patient is moved between units or hospitals, even if being treated for the same event. Hence, on some occasions, one event in the WOSCOPS system is matched to a number of events on the hospital discharge database acquired by linkage. The SRL database contains a stay marker which indicates when records relate to the same episode of hospital stay. Comparisons with the WOSCOPS database were carried out using the first admission and the last discharge dates of the continuous stay in a particular hospital rather than the actual dates appearing on individual SMR1 records, so that separate comparisons are made for transfers between hospitals, but not for transfers between units or consultants within a hospital. An additional problem was that the WOSCOPS system, being based on patient recall, often contained only approximations to the dates of events.

Each pair of files to be compared (Registrar General deaths and WOSCOPS deaths; SMR6 and WOSCOPS cancers; SMR1 and WOSCOPS hospitalisations) were joined on patient ID number (a unique number for each patient), and each combination of record pairs for a patient were compared (for deaths there should obviously only be one record pair for each subject). The comparison rules used were as follows:

1) agreement on patient ID number (always true when the patient appears in both files as record pair comparisons are only made within patient groups)

and some additional rules dependent on the record type.

For deaths:
2) exact agreement on date of death.

For cancer registrations:
2) agreement on year treatment commenced
3) agreement on month treatment commenced (assuming year is the same)
4) agreement on day treatment commenced (assuming month and year are the same).
Some flexibility on the date of treatment commencing was required since this date was not clearly defined on the WOSCOPS independent reporting system. Records were considered matched if any of the components of the date agreed.

For hospitalisations:
2) date of admission or date of discharge on the WOSCOPS record contained within the period of the SMR1 continuous inpatient stay.
For the SMR1 record the date of admission relates to the first admission of the continuous inpatient stay and the date of discharge relates to the last discharge of the continuous inpatient stay. The variation allowed on dates was necessary because of the inaccuracy of patient recall in reporting these dates. Since WOSCOPS adverse event reporting relied primarily on patient recall, it's dates were known to be slightly inaccurate at times.

Due to the relatively small numbers of events involved, and the importance of cardiac events to the WOSCOPS study, all cardiac surgery registrations and psychiatric hospital admission records obtained via record linkage were reviewed by a clinical research associate, who made a subjective decision as to whether or not they matched with a record from the WOSCOPS individual subject follow-up.

5.2.2 Checking of comparison results

For deaths, all of the computer matched records were checked by a clinical research associate to verify that the cause of death agreed on both records. For hospitalisation and cancer registration records, a check, by the clinical research associate, on the computer matches ensured that the diagnoses were in agreement. If there was any discrepancy between the two record sources then support documentation such as death certificates, histology reports and hospital discharge summaries were examined to confirm the true diagnosis.

Any linkage derived records which did not match up with a WOSCOPS record were investigated by a clinical research associate to verify that they did in fact relate to the WOSCOPS subject. Since the 6595 WOSCOPS randomised subjects were flagged with the National Health Service Central Register (NHSCR) for deaths and incident cancers, it was likely that notification for most of these events would eventually arrive from this source, and these events would then be documented by WOSCOPS. However, due to problems with the NHSCR system, some incident cancers are not notified, and many are notified late. The first stage of the investigation for SMR1s, SMR4s, SMR6s and SMR20s was a full review of the patient's file since it was sometimes possible to match the linkage SMR record with a WOSCOPS event which had dates differing by more than the limit for computer matching, or to find a reference to the event on another form in the patient's file. Sometimes the admission constituted an extra treatment for a known diagnosis, for example, an extra chemotherapy session for a known cancer. The second stage was to obtain copies of hospital discharge summaries and histology reports from

the appropriate consultants for any events still outstanding. These medical records enabled decisions to be made concerning whether the SMR records really belonged to the WOSCOPS subject or not.

Conversely, any WOSCOPS records which failed to match with an SMR record were examined to find a possible explanation. Possible explanations included stays in private hospitals (which are not required to submit SMR forms, although some do so voluntarily), admissions to hospices, or events that took place outside Scotland. Additional cancer records were searched for at the central Cancer Registry at ISD and at the West of Scotland Cancer Registry if necessary.

5.2.3 Documentation of new events

Any adverse event which was identified through record linkage but was not already recorded as part of the WOSCOPS independent follow-up was documented to enhance the validity of the final adverse event reporting of the WOSCOPS clinical trial. Since these events had not been reported to WOSCOPS trial centre staff by the patients in question, these events had to be documented at the Data Centre at Glasgow University and stored separately from the 'routine' trial documentation. The main reason for this was to prevent the staff at the trial centres from discussing these events with the patient, since it was assumed that the subject had deliberately not reported the event. Each of these events was investigated thoroughly by a clinical research associate, and hospital case notes and discharge summaries obtained where appropriate. These events were documented on adverse event forms similar to the routine forms described in section 2.6 and given in Appendix B, but marked with a blue stripe (see Appendix E for an example) to indicate that these forms were for Data Centre use only and should be kept separate. These forms were coded by the Adverse Event and Endpoints Committees where appropriate.

As the end of the WOSCOPS trial approached, additional linkages were carried out to interim files held at ISD to allow documentation of events occurring in 1994 or 1995 which had been missed by WOSCOPS individual follow-up. These updated linkages were not included in this comparative study since ISD did not yet consider these interim files to be complete, and it could not be guaranteed that WOSCOPS clinical researchers always completed 'routine' documentation if an event which they knew to have been recorded on 'blue-striped' forms then came to light through normal channels. The results reported below thus consider only linkages to the main SRL databases.

5.3 Results

The results of the validation exercise are considered separately for each of the different databases involved in the linkage exercise - Registrar General deaths, cancer registrations (SMR6s), non-psychiatric hospitalisations (SMR1s), psychiatric hospitalisations (SMR4s), and cardiac surgery registrations (SMR20s) (see section 3.4.1 for a description of these databases).

5.3.1 Deaths

At the time of linkage, the WOSCOPS database contained records of 165 deaths until the end of 1993. The computerised record linkage to the Registrar General (RG) Death records identified 166 deaths for the WOSCOPS subjects. 164 of these records were common to both files, with exact agreement on the dates of death. The causes of death identified by the two approaches for these 164 deaths were in agreement. There was an exact match on RG and WOSCOPS cause of death codes in 154 cases and a minor disagreement on the coding of cancer sites for 6 deaths. The remaining cases were as follows,

1) the WOSCOPS database contained the primary (bronchopneumonia) and secondary (disease of the endocrine glands) causes of death as stated on the death certificate (obtained from the Registrar General for Scotland), while the SRL database had reversed the status of the causes of death.
2) the WOSCOPS database identified the cause of death as pulmonary embolism, confirmed by the death certificate and hospital notes. The SRL database stated the cause of death as malignant neoplasm of the trachea, bronchus and lung. The patient was known to suffer from cancer of the bronchus, but this information was not stated on the death certificate.
3) the WOSCOPS database contained only the primary cause of death (myocardial infarction). The death certificate also had the secondary cause of death of hyperlipidaemia. The SRL database had reversed the status of these two causes of death on the death certificate.
4) the WOSCOPS database identified the cause of death as smoke inhalation while the SRL database stated the cause of death to be abuse of drugs. The original death certificate gave the cause of death as inhalation of toxic fumes in a house fire, while a revised death certificate changed this to asphyxiation due to alcohol abuse.
WOSCOPS death data is confirmed from death certificates, post mortem examinations, hospital discharge summaries and reports from GPs when available, but the WOSCOPS Endpoints Committee is not always in agreement with the death certificate.

The one death identified by WOSCOPS follow-up but not by record linkage had occurred outside Scotland and thus did not appear on the Registrar General for Scotland's file of deaths. This missed death was not due to any error in the linkage process.

Of the 2 deaths identified by record linkage but not by WOSCOPS independent follow-up, 1 was an internal linkage error on the SRL database which matched 2 deaths with the same subject. This type of error is corrected routinely by clerical checking at SRL and would not have affected the WOSCOPS linkage if the linkage exercise had been deferred until SRL's internal checking had been completed. The other death was correct

for the WOSCOPS subject who had been assumed to be defaulting visits, and there had been no response to attempts at follow-up. Flagging with NHSCR had also failed to alert WOSCOPS staff to this death by the time of this comparison, over 2 years after the subject's death.

These results are summarised in Table 5.1.

		Identified by CRL		
		Yes	No	Totals
Identified by WOSCOPS	Yes	164	1	165
	No	1	?	1
	Totals	165	1	166

Table 5.1 True deaths for the WOSCOPS subjects identified by either or both follow-up systems.
? denotes the unknown number of deaths missed by both follow-up methods.

5.3.2 Cancer registrations

Up until the end of 1992 the computerised record linkage and the WOSCOPS system identified 137 and 142 primary malignant neoplasms respectively. 131 cancers were common to both systems, of which 52 agreed completely on the date of treatment commencing, and a further 43 had the date of treatment commencing occurring in the same month.

6 events on the SRL system were not identified by the WOSCOPS system, of which 5 were correct for the WOSCOPS subjects. 3 related to regular attendees and were for a malignant melanoma removed by the GP and 2 basal cell carcinomas removed in hospital. Two other events (a basal cell carcinoma and a prostatic carcinoma) not

identified by WOSCOPS were related to subjects who had withdrawn from trial medication. The further additional SRL record did not relate to a WOSCOPS subject. This false positive linkage was due to internal linkage errors at ISD, with this record not achieving the threshold criteria on its own but being linked along with a higher scoring record in the same patient group.

Of the 11 unmatched cancer registrations on the WOSCOPS system, 2 events were not picked up on the SRL linkage because of errors in patient identifiers in the SRL records (1 where the forename and surname were transposed, and the other where the subject used his middle name and there was an error on date of birth), and 9 events (6 of which were skin cancers) had not yet been reported to the Scottish Cancer Registry. With regard to the 6 skin cancers, basal and squamous cell skin carcinomas have always caused documentary problems since some physicians would not classify them as true cancers. Skin cancers are often dealt with by GPs and are not referred to hospital. This can result in a failure to register these cancers, although they should be. On account of this problem of poor reporting, ICD9 (International Classification of Diseases, 9th revision) code 173 (other malignant neoplasm of skin) is routinely excluded from national statistics.

These results are illustrated in Table 5.2 for cancers that truly belong to the WOSCOPS subjects.

		Identified by CRL		
		Yes	No	Totals
Identified by WOSCOPS	Yes	131	11	142
	No	5	?	5
	Totals	136	11	147

Table 5.2 True cancer registrations identified for the WOSCOPS subjects by either or both systems of follow-up. ? denotes the unknown number of events missed by both follow-up methods

On investigation of the support documentation for cancer registrations held by WOSCOPS, it was found that several events had been documented by WOSCOPS as cancers, in addition to the 142 primary malignant neoplasms investigated above, which were in fact benign (2 basal cell skin papillomas, 2 pituitary adenomas and 1 parotid adenoma), and were therefore not registered on an SMR6 form and should not have been documented by WOSCOPS. 7 additional WOSCOPS cancer registrations were not matched with an SMR6 because they related to recurrences of cancers for which an SMR6 had previously been completed when the cancer originally occurred, prior to the subject's randomisation. WOSCOPS could only document events which had occurred since the subject was randomised into the study, and thus considered any cancer recurring after randomisation as a new cancer. These events which were not genuine new cancers have been omitted from these results as, strictly speaking, they should not have appeared in either database.

In summary, of the 147 separate non-benign, post-randomisation primary neoplasms relating to a WOSCOPS subject which were identified by one or other system, 136 (93%) were picked up by SRL and 142 (96%) were eventually identified by the WOSCOPS system.

5.3.3 Non-psychiatric hospitalisations

The WOSCOPS database contained 2791 records for non-psychiatric, non day case discharges occurring before the end of 1993, and the computerised linkage generated 3403 records which related to 2952 episodes of continuous inpatient stay. These datasets contained 2606 common hospitalisations. There were 2931 SMR1 records relating to these 2606 episodes of hospital stay. Thus, 325 of the additional SMR1 records corresponded to hospitalisations which were double-reported or involved transfer of the patient between wards or hospitals during an event which had been identified within the WOSCOPS system. Of these 2606 episodes of stay, 1532 (59% of the 2606 matches) agreed exactly on both date of admission and date of discharge on the WOSCOPS

records and SMR1 stays, while 570 agreed exactly on one of these dates. A further 127 were only one day out on either date of admission or date of discharge, while 23 were two days out on either date of admission or date of discharge. The remaining 354 SMR1 records (including events reported late on the WOSCOPS system) were matched up clerically by a clinical research associate.

The additional 185 records in the WOSCOPS database were made up of 33 events which took place outwith Scotland, 94 events which were associated with private hospitals or hospices and 58 events which did not appear on the SRL database and for which no explanation for their absence could be identified. This group of 58 events would be expected to be associated with failures in the record linkage system due to errors in the key identifiers in one or other of the systems, or to events for which appropriate notification forms were not completed in the hospitals involved. These results for unmatched WOSCOPS records are given in Table 5.3.

Resolution	Number of records
Out of Scotland	33
Private hospitals / hospices	94
Missed by linkage	58
Total	185

Table 5.3 Resolution of WOSCOPS non-psychiatric hospitalisation records not matched with an SMR1

For computerised record linkage, 71 of the 472 unmatched records were additional treatments for known events (for example, additional chemotherapy sessions for a known cancer). Since the source events were identified in the WOSCOPS system, these additional treatments were not considered in this analysis to represent errors in the WOSCOPS system. 76 hospitalisations corresponded to adverse events which had been reported within the WOSCOPS system but had not been identified as a hospitalisation

due to failures in the WOSCOPS documentation procedures, but these events may have been detected in a review of patient files. 114 events were clearly linkage errors and had incorrectly been associated with WOSCOPS subjects. 210 events were confirmed as being previously unreported hospitalisations for WOSCOPS subjects. One more SMR1 record could not be verified as to whether or not it related to a WOSCOPS subject, due to unavailability of the hospital case notes. These results for unmatched SMR1 records are summarised in Table 5.4.

Resolution	Number of records
Events not coded as a hospitalisation	76
Additional treatments	71
New Events	210
Could not be verified	1
Linkage errors	114
Total	472

Table 5.4 Resolution of SMR1 records not matched with a WOSCOPS non-psychiatric hospitalisation

The 3077 events relating to WOSCOPS subjects identified by either system (ignoring double reported events, transfers and extra treatments for known events) are summarised in Table 5.5.

		Identified by CRL		
		Yes	No	Totals
Identified by WOSCOPS	Yes	2606	185	2791
	No	286	?	286
	Totals	2892	185	3077

Table 5.5 Distribution of true hospitalisations identified for the WOSCOPS subjects by either or both systems of follow-up. ? denotes the unknown number of events missed by both follow-up methods.

2791 (91%) were eventually identified within the WOSCOPS system and 2892 (94%) were identified by computerised record linkage. In addition, the record linkage approach identified one record which could not be confirmed as belonging to a WOSCOPS subject and wrongly associated 114 events with WOSCOPS subjects. In fact, these 114 events related to only 42 subjects.

Calculation of the standard measures of sensitivity and specificity for the two approaches is difficult because there is no gold standard method which guarantees to identify all true events (as emphasised by the question mark in Table 5.5) and because it is the number of events that is being dealt with and not the number of subjects. However, if it is assumed that almost all events have been picked up by one or other of the methods then the WOSCOPS and record linkage systems will clearly have high sensitivities as evidenced by the 91% and 94% success rates in identifying events. In terms of specificity, the WOSCOPS system requires support documentation for all events and relies on patient reporting, so it could be considered to have a zero error rate in terms of wrongly identifying a subject as having had an event. In the record linkage approach only 114 events (on 42 subjects) were incorrectly identified, suggesting that the system would have a high specificity.

Of the 210 new events which were not identified by the WOSCOPS system 143 were associated with subjects who had either formally withdrawn from taking study medication or were defaulting from attending at routine study visits. For this purpose a defaulter was taken to be a participant who has missed at least three trial visits immediately following the date of event. The remaining 67 were associated with subjects who were regularly attending trial visits, that is, they had not withdrawn and were not defaulters. See Table 5.6 for further details of the default status of the subjects with events unreported to WOSCOPS.

Resolution category	# records	# subjects
Withdrawn before event	123	86
Temporarily defaulting at time of event	20	16
Regular attendee (around time of events)	67	60
TOTAL	210	162

Table 5.6 Default status of subjects with SMR1 events not reported to WOSCOPS.

The reasons for these 210 hospitalisations were categorised by a clinical research associate as in Table 5.7. It is interesting that 30 of the 210 events were cardiac in nature and hence very important for the WOSCOPS study. In fact, three of these, for myocardial infarctions, turned out to be undocumented primary trial endpoints. In the regular attendee group there was a relatively high number of events in categories which patients might be more likely to conceal from trial staff (for example, urogenital problems and psychological problems - which included 5 overdoses).

Diagnosis	Withdrawals and defaulters	Regular attenders	Total
Urogenital	20	12	32
Cardiac	22	8	30
Gastrointestinal	19	5	24
Orthopaedic	15	7	22
Respiratory	9	5	14
ENT	12	2	14
Vascular	10	4	14
Psychological	5	7	12
Eye related	3	7	10
Neurological	5	4	9
Cancer	7	2	9
Skin related	4	3	7
Hernia	5	1	6
Dental	2	0	2
Endocrine	1	0	1
Miscellaneous	4	0	4
Total	**143**	**67**	**210**

Table 5.7 Diagnoses of events not reported to WOSCOPS. The numbers relate to the number of SMR1 records which formed new adverse events and were documented on blue-striped forms, for each diagnosis category.

5.3.4 Other uses of SMR1 information

In addition to the identification of previously undocumented events the linked computer records potentially provide additional information. The SMR1 records can provide additional flags of potentially serious events, such as the type of admission involved, or whether the event required transfers between units (such as ITU), consultants,

specialties (medical or surgical) or hospitals. They can also be used as support documentation for adverse events which have been documented by the WOSCOPS follow-up system.

Type of admission

The SMR1 records contain a marker indicating the 'type of admission'. The possible types of admission are:

A) deferred admission

B) waiting list admission

C) repeat admission

D) transfer

E) emergency admission for deliberate self-inflicted injury or poisoning

F) emergency admission for road accidents

G) emergency admission for an accident in the home

H) emergency admission for other injury

I) other emergency admission

Since accidents and suicide attempts may be important events for trials of cholesterol lowering drugs, these admission codes provide a useful flag on these serious events and an additional means of verifying how they have been coded on the WOSCOPS database. All SMR1 emergency admissions of types E)-H) (that is, for 'trauma') were cross-checked with WOSCOPS records and hospital discharge summaries by a clinical research associate. Most of the 121 SMR1 emergency events identified were in reasonable agreement with the WOSCOPS database, but for two events, the diagnosis on the WOSCOPS database was recoded after closer examination of the hospital case notes:

1. from 'depression' to 'drug overdose'
2. from 'abuse alcohol' to 'drug overdose'

Transfers

Separate SMR1 records are completed for all transfers between hospitals, units (such as intensive care), consultants and specialties (for example, general medicine, orthopaedic surgery or neurology, although the consultant will usually change with the specialty). Transfers will not usually be required for a 'straightforward' admission, so they can be used to identify potentially more serious events. Where there were more than 2 SMR1 records relating to a single episode of continuous inpatient stay in hospital, all documentation for these events was reviewed by a clinical research associate. This minimised the possibility of missing a serious event which occurred during an admission for another reason.

Up until the end of 1993, there were 45 hospital stays involving 3 SMR1 records each, 10 involving 4 SMR1 records each, 2 involving 5 records each, and 1 hospital stay which involved 6 SMR1 records. These were all reviewed in detail by a clinical research associate. Of the 191 individual SMR1 records making up these 58 episodes of stay, 52 were admissions from home, 6 were admissions from another hospital and 133 were admissions from another unit within the same hospital. Correspondingly, 51 records related to discharges to home, 4 to discharges to another hospital, 133 to discharges to another unit within the same hospital and 3 to patient deaths.

Cardiac day case admissions

An SMR1 record is completed for every admission to hospital, including day case admissions. Although day cases were not documented as hospitalisations in the WOSCOPS study, it was important to obtain information on all cardiac events, which were documented as non-hospitalisation adverse events. Coronary arteriography (a secondary endpoint on the WOSCOPS study) is now often carried out on a day case basis.

The completion of SMR1 records for day cases allowed a review of day admissions for cardiac procedures by a clinical research associate, as a further check on the completeness of the WOSCOPS independent follow-up. There were 45 day case SMR1

records for cardiac procedures which related to the WOSCOPS randomised subjects up until the end of 1993. 40 of these had already been documented by WOSCOPS. Of the remaining 5, 2 were referred to on other forms but the correct documentation had not been completed, 1 related to a 'failed' angiogram and did not require documentation, and 2 were events previously unknown to WOSCOPS.

Support documentation

WOSCOPS routinely obtains support documentation (for example, discharge summaries or hospital case notes) for all Trial Endpoints (deaths and serious cardiac events) and cancers, but not for every hospitalisation. Where a WOSCOPS hospitalisation record has been matched with an SMR1 record, the SMR1 record may be used as support documentation for the purposes of verifying that the hospitalisation occurred, since it isn't feasible to obtain hospital discharge summaries for all hospitalisations. The difference in dates between some of the SMR1s and WOSCOPS records is not a new problem, as the dates on hospital discharge letters currently used as support documentation often differ from the dates on the WOSCOPS records (which are based on patient recall).

5.3.5 Psychiatric hospitalisations

Since there was a very small number of psychiatric hospitalisations (documented on an SMR4 form, see Appendix C), they were checked clerically by a clinical research associate who made a subjective decision based on his experience, as to whether they related to a WOSCOPS subject.

The computerised linkage identified 41 non-day case, inpatient psychiatric hospitalisations until the end of 1993. 30 had already been identified by the WOSCOPS system, 7 were previously unknown events, and 4 were record linkage errors and did not relate to a WOSCOPS subject. Of the 7 new events discovered, 3 were for depression, 2 relating to a regular attendee and 1 to a subject who had withdrawn from trial

medication, while the other 4 were for alcohol dependence, 2 relating to regular attendees and 2 relating to subjects who had withdrawn from trial medication.

The WOSCOPS follow-up had identified 31 psychiatric hospitalisations until the end of 1993. 28 of these matched with the 30 SMR4 records. 1 of the 3 records not found by linkage related to an ongoing, longstay admission for depression. Of the other 2, 1 was for depression and the other for alcohol abuse. These results are summarised in Table 5.8.

		Identified by CRL		
		Yes	No	Totals
Identified by WOSCOPS	Yes	28	3	31
	No	7	?	7
	Totals	35	3	38

Table 5.8 Distribution of true psychiatric hospitalisations identified for the WOSCOPS subjects by either or both systems of follow-up. ? denotes the unknown number of events missed by both follow-up methods.

5.3.6 Cardiac surgery registrations

While the previous linkages have all involved linked databases, the SMR20 data is held 'on-line'. This means that the SMR20 register being linked to is completely up to date (rather than only up until the end of the previous year), and that the SMR20 data is held as single records, not in patient groups. The SMR20 file also contains records for people who are on the waiting list for cardiac surgery but have not actually been admitted yet.

Since relatively few SMR20 records are received each year, and cardiac surgery is an important adverse event in the WOSCOPS study, all of the SMR20 records were checked clerically by a clinical research associate.

68 events were identified by record linkage until April 1994 (when this linkage was carried out), 64 of which had already been documented in the WOSCOPS system or were in the process of being documented, and 4 which were linkage errors and did not relate to a WOSCOPS subject. It was reassuring to note that WOSCOPS independent follow-up did not appear to have missed any of these major cardiac events.

5.4 Influences on the quality of linkage

An assessment was made of the effects on the quality of linkage results of
1) using the patient identifying information collected at the first screening visit and entered directly into a database by screening centre nurses as opposed to updated data, prepped from forms by clerical staff
2) using unpostcoded data (as the WOSCOPS data would have been without the postcoding exercise carried out).

This assessment was carried out for death linkages only, as linkages to the incidence databases would have resulted in a lot of extra clerical checking work to be done by the clinical research associate, while verification of a subject's mortality status was relatively simple to ascertain. These assessments have used linked death data up until the end of 1993.

5.4.1 Use of linkage information collected at first screening visit

Death linkage was carried out for the 97,165 WOSCOPS screenees with sufficient identifying information in a similar way to, but completely separately from, the linkage for the 6595 randomised subjects as described in chapter 4. For the trial linkage, linking information was updated by 'duplicate' records, while for the screenee linkage, only a single baseline record obtained at screening visit 1 was available. The links identified for the 6595 WOSCOPS randomised subjects were extracted from the screening linked files, and these links were compared to the links achieved in the routine trial linkage in order to investigate the effects of using screening visit 1 data as opposed to data updated throughout the trial. The question was whether it was really worth the time and effort of updating patient identifying information, in terms of improved linkage quality. (It should be noted that not only does screening visit 1 data contain a single address, but it was entered directly onto the computer by screening centre nurses as the patient visit took place, while the trial data was entered on forms and then processed by experienced data clerks. This may be a factor in the quality of the data collected in addition to the opportunity to match being restricted to a single postcode, for screening visit 1 data, when the subject may have moved house.)

Until the end of 1993, routine linkage of the 6595 randomised subjects identified 166 deaths, while screening linkage identified 165 deaths for these subjects. 157 of these deaths were common to both linkages (trial and screenees), with 155 of these being in exact agreement with the WOSCOPS independent database of deaths. One common link was due to an ISD internal linkage error, with 2 deaths occurring for the same subject. This false link to a WOSCOPS subject would not have occurred if ISD's routine clerical checking program had been completed before the WOSCOPS linkages were carried out. The other death common to both linkages had been missed by WOSCOPS trial centre staff.

9 deaths were found by the routine trial linkage but missed by the screenee linkage. These deaths were in agreement with the WOSCOPS independent database. 9 missed deaths out of 165 deaths identified correctly for the WOSCOPS randomised subjects by routine linkage gives a false negative error rate (where genuine deaths for the WOSCOPS subjects were missed by the screenee linkage but not by the trial linkage) of 5.4% of deaths, for the screening death linkage relative to the trial linkage.

Conversely, 8 deaths were identified by the screening linkage but missed by the routine trial linkage. All of these subjects have attended trial visits since their proposed date of death (some proposed deaths were even prior to randomisation). 7 out of these 8 links were admitted because the threshold score set for the screening linkage was lower than for the trial linkage (26 instead of 30 - the distribution of scores varies with differing quality of linkage information and thus the threshold score must be adjusted according to the characteristics of the datasets). However the other death had a linking score of 46, which could only be achieved by almost exact agreement on all identifying information (but the death could not have related to the WOSCOPS subject since the date of death was prior to his randomisation). So, out of the 165 deaths identified by the screening linkage, there was a false positive error rate of 4.8%.

The increased error rates for the screenee linkages relative to the routine trial linkages are likely to be greatest for the death linkages, since earlier records in patient groups may provide better postcode matches to the screening visit 1 data and thus link the whole patient record group for 'incidence databases', while the death record will have only a subject's final address which may well be different from the address at screening. The potentially increased error rates resulting from use of screening visit 1 data rather than updated data will increase as the length of time since screening increases, but for a mean follow-up of 3.67 years (from randomisation date until the end of 1993), the linkage quality seems fairly reasonable, although not as good as the linkage using updated trial data which yielded virtual 100% accuracy, as described in section 5.3.1.

5.4.2 Lack of postcodes

Death linkage was carried out for the 6595 WOSCOPS randomised subjects using identifying information without the postcodes, completely independently of the routine WOSCOPS trial linkage using baseline and updated postcodes.
The links achieved with and without postcode information were then compared, in order to assess the benefit arising from the WOSCOPS postcoding exercises, and thus whether or not it is worth postcoding addresses in general, prior to carrying out linkage exercises.

Routine trial linkage identified 166 deaths, while unpostcoded linkage identified 163 deaths for the WOSCOPS randomised subjects.

159 death records were common to both linkages:

- 157 which were in exact agreement with the WOSCOPS independent database of deaths
- 1 internal linkage error - there were 2 deaths for the same subject
- 1 death which had been missed by WOSCOPS

7 deaths were found by the routine trial linkage but missed by the unpostcoded linkage - these deaths were all in agreement with the WOSCOPS independent database. These 7 missed links are linkage errors due to the increased difficulty in setting a linkage threshold when the postcode information is unavailable. 7 missed deaths out of 165 deaths identified correctly for the WOSCOPS randomised subjects gives a false negative rate of 4.2% relative to the postcoded trial linkage.

4 deaths were identified by the unpostcoded linkage but not by the routine trial linkage for WOSCOPS randomised subjects. All of these subjects have attended trial visits since their proposed date of death (2 proposed dates of death were even prior to the randomisation date for the subject in question). So out of the 163 deaths identified by the unpostcoded linkage there was a false positive error rate of 2.4%. These 4 deaths were included in the unpostcoded linkage as the threshold score for this linkage was

only 25, while it was 30 for the routine trial linkage with postcodes. However, without postcode information, all of the linkage scores were lower and many more correct links would have been missed if the threshold had not been lowered. In fact, only 88 out of the 163 links made using the unpostcoded data achieved a linkage score greater than or equal to 30.

The differences in the score frequencies for the postcoded and unpostcoded linkages are shown in Figure 5.1. As can be seen from the graph, the threshold area is clearer for the postcoded linkage, with little to be lost by setting the threshold anywhere between 26 and 31, while for the unpostcoded linkage, the threshold region is narrower (between 23 and 25) but contains a higher frequency of record pairs at each score. The separation between 'good' and 'bad' matches is less distinctive, and thus the threshold is harder to set, for unpostcoded data. Where no postcodes are available, the subjectivity of the threshold setting becomes a more important factor.

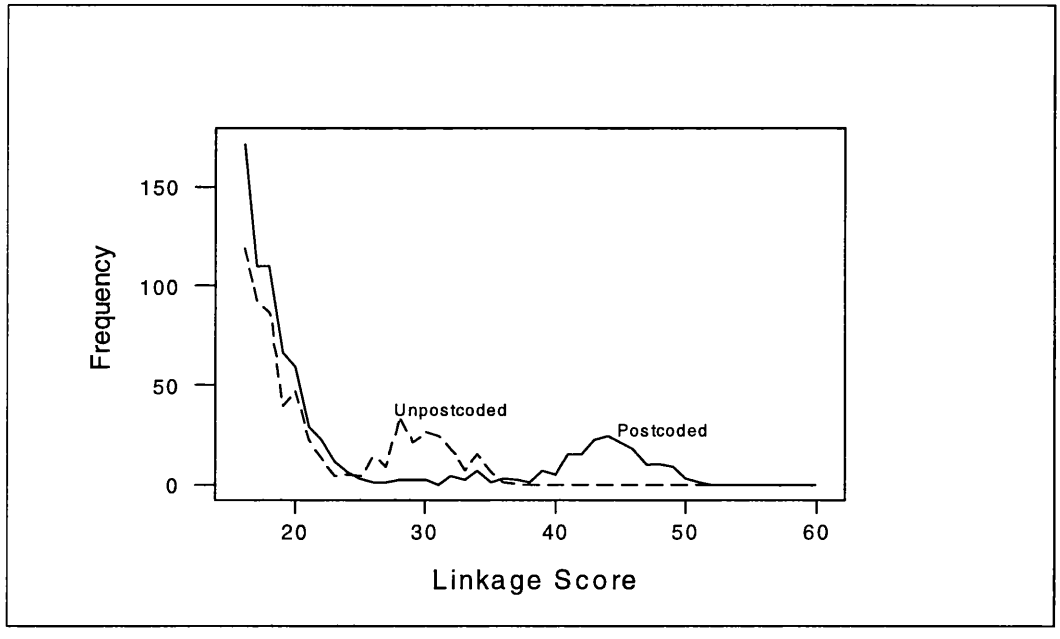


Figure 5.1 Score frequencies for postcoded and unpostcoded links

Figure 5.1 also shows that the basic distribution of scores for linked data is similar for both postcoded and unpostcoded linking information, the actual scores appear to be simply shifted downward when there are no postcodes, closer to the distribution of unlinkable pairs which remains virtually the same for both linkages, as can be seen from the lower end of the score frequency graph in Figure 5.2.

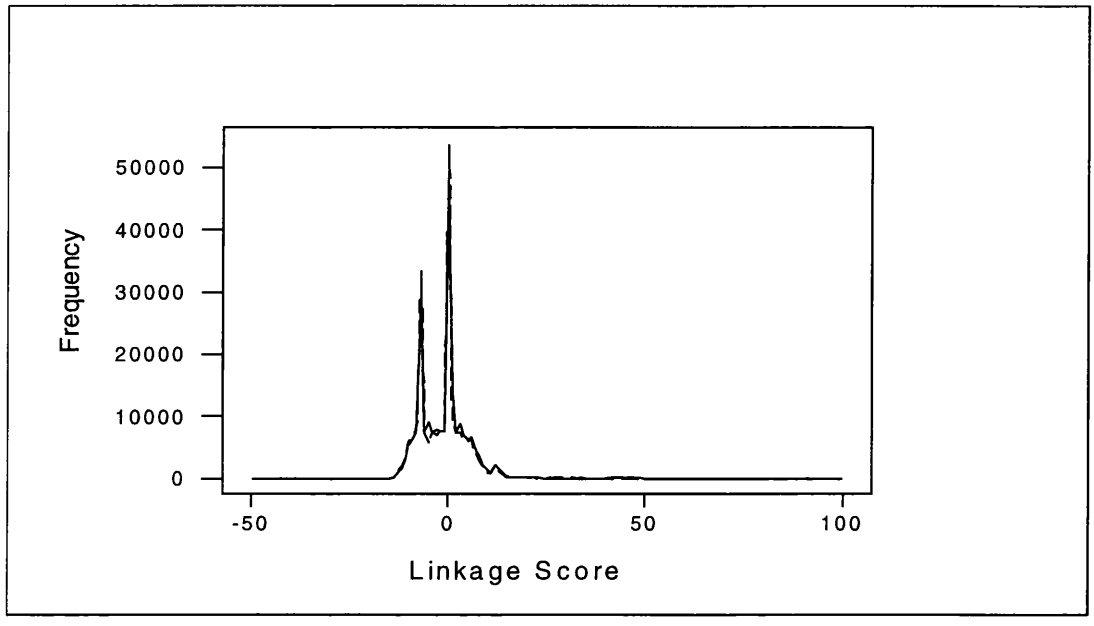


Figure 5.2 Score frequencies for postcoded and unpostcoded non-links

For the 158 deaths where a correct link was achieved by both methods (postcoded and unpostcoded), the paired scores were plotted against each other (Figure 5.3) and revealed a clear relationship between the two scores, with a shift upwards where postcodes were available (in agreement with the frequency graph in Figure 5.1) for most of the linked records. However, this shift does not occur for all records, and some additional discriminating power can be derived from use of postcodes. It should be noted that scores can only be plotted when they appear in both linked files. Records which failed to achieve the threshold score are not available for analysis and thus cannot be positioned on the graph.

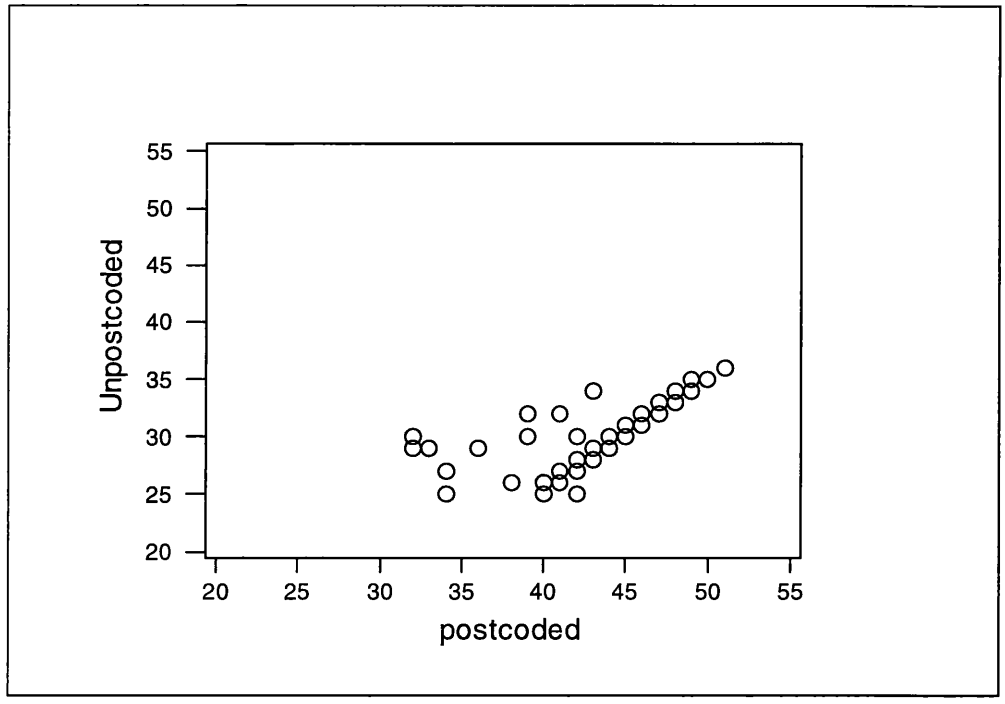


Figure 5.3 Unpostcoded score against postcoded score for the 158 deaths where a link was achieved by both linkage methods. These scores can only be plotted in this way when they appear in both linked files, since records not achieving the threshold score are not available for analysis and thus cannot be positioned on the graph. The threshold score was 30 for the postcoded linkage and 25 for the unpostcoded linkage.

For deaths until the end of 1993, unpostcoded linkage information provides a reasonable quality of linkage. It is expected that potential errors resulting from lower quality linkage information will be more frequent for the hospitalisation (SMR1) linkage than for the death linkage investigated here as death records contain full forenames rather than just first initial and the data are generally of a higher quality on a death certificate.

5.5 Discussion

The validation of the linkage to the Registrar General's Deaths database was very encouraging, with a virtual 100% match being achieved in terms of the deaths being identified by both linkage and routine WOSCOPS follow-up. The WOSCOPS system involved manual flagging with the Registrar General for Scotland's death register in addition to individual subject follow-up, so it would be expected that these events would be reliably reported in the WOSCOPS system. This comparison for deaths was primarily an assessment of the computerised matching of linkage information. It is encouraging that this yielded no false positives (except for the miss-link due to failing to wait for completion of SRL's checking procedures before carrying out the linkage) despite the fact that the linkage was based only on name, sex, date of birth and postcode.

The information obtained from the linkage on cause of death is based on death certificates. Concern has been expressed about the use of death certificate data (Alderson et al, 1983). However, others have concluded that such data are valid epidemiologically for most cancers and for cardiovascular and coronary heart disease deaths. A previous study in Scotland (Isles et al, 1986) showed that notification from the Registrar General was virtually complete and a comparison of coding of causes of death by independent physicians yielded no important discrepancies. In clinical trials, fatal events identified by record linkage could be fully investigated to check on the accuracy of the coding of cause of death, but in large epidemiological studies this would not be a feasible option. However, it would appear that studies based on broad classes of events are not subject to substantial bias.

For the hospitalisation linkage there was a much better opportunity to study the frequency of both false negatives and false positives. Although the SRL linkage had an overall false negative rate of 185 in 3077 (6%) events identified by either system, 127 of these events occurred either outside Scotland or in private hospitals or hospices and were unlikely to be identified by the linkage system. At most 58 (1.9%) of the events

could have been missed because of linkage mismatches, although many of these were probably due to the failure of the hospital concerned to report the event. This was achieved at the expense of 114 false positives relative to the 3077 events for WOSCOPS subjects identified by either system. Of these 114 false positive records, 38 individual records did not achieve the threshold score and were only linked because of the practice of linking a whole group of patient records on the basis of the highest score achieved by any record in that group. These rates compare well with the false negative and false positive rates (2%-3%) which SRL staff predicted on the basis of their own unpublished investigations. The current study provides an important independent check on the quality of the linkage carried out at SRL. The quoted error rates are based on 'clinical' events and would be lower if transfers and additional treatments were included. It should also be noted that the calculations assume that all events have been identified and that there are at most a very small number of events which have been missed by both methods of follow-up. The numbers of events which have currently been reported and cross-checked for the other database linkages are still rather low for any definitive conclusions to be drawn. With the exception of skin cancers which are typically underreported, the cancer linkage was encouraging and it was interesting that 7 of the 43 psychiatric hospitalisations relating to WOSCOPS subjects had not previously been identified by the WOSCOPS follow-up. It would appear that the 'smaller' linkages (psychiatric hospitalisation and cardiac surgery) may be less accurate than the linkages to larger databases. This may simply be a superficial result due to the smaller numbers of events involved. The fact that approximately two thirds of the previously unidentified hospitalisations were associated with men who had either withdrawn or were defaulting from trial visits confirmed preconceptions that event rates might be underreported in this category of subjects. It is particularly worthy of note that the SRL computerised linkage alone compared very favourably with the active patient follow-up system employed in WOSCOPS.

These results can be compared to the results obtained using linkages up until the end of 1991 and reported by the West of Scotland Coronary Prevention Study Group (in press). These results may be considered to be of a potentially higher quality than the later

comparisons since they are uninfluenced by the effects of sorting out potential double documentations (that is, on both routine and blue-striped forms) which became a problem towards the end of the trial. It is reassuring to note that the results obtained in the comparison study reported in this chapter are similar to those found in 1991. For example, for non-psychiatric hospitalisations until the end of 1991 the results are summarised in Table 5.9. Of the 1246 events relating to WOSCOPS subjects, 89% were eventually identified within the WOSCOPS system and 95% were identified by computerised record linkage. In addition, record linkage wrongly associated 52 events with the WOSCOPS subjects. These proportions are clearly similar to those obtained in the later comparisons reported above. Further details of the 1991 results can be found in West of Scotland Coronary Prevention Study Group (in press).

		Identified by CRL		
		Yes	No	Totals
Identified by	Yes	1043	66	1109
WOSCOPS	No	137	?	137
	Totals	1180	66	1246

Table 5.9 Distribution of true hospitalisations identified for the WOSCOPS subjects by either or both systems of follow-up until the end of 1991. ? denotes the unknown number of events missed by both follow-up methods.

In summary, this comparative study demonstrated minor flaws in both systems of subject follow-up and showed that follow-up based on computerised linkage alone can be as effective as reporting based on direct contact with the patients. Active patient follow-up avoids the dangers of computerised linkage errors but suffers from poor follow-up for withdrawals and defaulters, and may under-report events for 'sensitive' diagnoses. Computerised record linkage, on the other hand, is relatively fast and cheap and is good for picking up events for withdrawals and defaulters. However, it only covers events occurring in Scotland and reporting is incomplete from private hospitals.

Based on our experience, in contexts where computerised follow-up is feasible (such as a community based primary prevention study), such a system provides not only significant cost advantages but also potentially more complete follow-up of serious adverse events. It may thus be used to enhance the validity of the reporting of serious adverse events in clinical trials. What cannot be questioned is the fact that computerised linkage systems can significantly enhance the completeness of patient follow-up obtained by more traditional methods in clinical trials, both by providing independent validation of trial data and by providing additional data. Recently, the issue of fraud in industry and government funded clinical trials has attracted a great deal of media attention both in Europe and in North America. Validation of major trial outcome measures against independently held databases using record linkage techniques is one way of partially addressing this problem. If suitably validated, these systems also provide invaluable resources for epidemiological studies.

Previous reports of similar comparison exercises have had mixed success. A previous study in Scotland (Hole et al, 1981) reported a validation of computerised record linkage between an external data set and records held at ISD against manual tagging of subjects using the National Health Service Central Register (NHSCR) using the cohort of 3062 patients involved in the Renfrew and Paisley Study. This linkage was carried out prior to the creation of the linked database and thus used different procedures to those currently employed at ISD. This study obtained sensitivities of 66% and 67% for links to mortality and inpatient hospitalisations respectively, based on computerised record linkage alone, changing to 65% and 81% when clerical checking was added to their system. The difference between this earlier study and the current one could be due to improvements in the computerised record linkage system now in use at SRL, the quality of the data recorded on nationally held records and the quality of the recording of patient identifying information in the WOSCOPS trial.

A number of studies have been carried out in Canada, with the general finding that computerised record linkage has performed no worse than individual subject follow-up. In these Canadian studies, the computerised record linkage and the manual follow-up

both dealt with the same basic death file. Statistics Canada carried out a comparison of individual patient follow-up with computerised record linkage to the Canadian Mortality Database (Shannon et al, 1989). The study sought the mortality status in 1985 for 2469 employees, at an industrial plant in Ontario, who were alive in 1977. Where the 2 methods of follow-up gave a different outcome, further investigation was carried out and the individual patient follow-up was, surprisingly, found to be in error for most of these cases. It was thus found that computerised record linkage gave accurate information and was relatively cheap for large studies (>1000 cases suggested). It would be preferable to compare both methods of follow-up to the absolute true mortality status (unknown), but investigation was only carried out where the 2 methods gave a different outcome, and it is possible that both were mistaken in some cases. A similar study involving uranium workers in Eldorado (Newcombe et al, 1983) found that not only was the computer more successful at finding matches than the manual searchers, but it was also less likely to yield false linkages with death records not related to the study population.

The comparative study reported in this chapter has thus achieved results which compare well with previous studies. This may be due to the extensive 'cleaning' exercise which has been carried out on the WOSCOPS database, yielding a high quality of patient identifying information to be used in the record linkage process. The techniques employed in record linkage have improved over the years since the previous Scottish study was carried out and this may therefore not be a fair comparison. However, in agreement with this study, the Canadian studies achieved highly accurate results for linkages to mortality databases. The main drawback of the Canadian system compared to the Scottish system used here was that the Canadian system involves only the mortality database, while the Scottish system makes use of all records which form part of the National Health Service databases. The Scottish system can thus give much more versatile and complete information for subjects to be followed up. On the whole, the Scottish Record Linkage System compares favourably with other linkage systems and is more accurate than the system available in the early 1980s.

Chapter 6

Analysis of linkage data for the WOSCOPS cohort

6.1 Introduction

The WOSCOPS screening exercise was of great interest in its own right in a part of the world where coronary heart disease death rates are consistently among the highest in the international league tables of cause specific mortality (see section 1.2.5).

Full informed consent was not available to link the WOSCOPS screened subjects to their computerised medical records on an individual case basis. However, permission was obtained from the Scottish Record Linkage System (SRL) to hold the identifying information for these subjects in Edinburgh, to link them to the SRL databases and to make tabulations of frequencies of events, defined by cross-classifications of risk factors collected during screening, available to WOSCOPS researchers. For instance, a table of the frequency of coronary heart disease death could be obtained broken down by smoking status and quintiles of age, total cholesterol and diastolic blood pressure. In this way, analyses of the relationships between outcome and the various risk factors were possible.

6.2 Methods

6.2.1 Data available

In the screening phase of WOSCOPS 105,383 subjects were screened for coronary heart disease risk factors. The information collected on each subject was described in chapter 2 and the detailed characteristics of these subjects have been reported elsewhere (section 2.5 and West of Scotland Coronary Prevention Study Group, 1995). For the 97,165 subjects who attended screening visit 1, and had sufficient identifying information to allow linkage to be carried out, data were available on demographic parameters and coronary heart disease risk factors (Table 2.1).

The postcoding exercise which was carried out (section 4.3) not only enhanced the quality of the linkage to SRL databases (section 4.4), but also allowed assignment of Carstairs deprivation scores (Carstairs and Morris, 1991). These deprivation scores were derived from the 1991 census data, and were assigned to WOSCOPS subjects on the basis of postcode sector. The Carstairs score is based on proportions of car ownership, male unemployment, low social class and overcrowding (Carstairs and Morris, 1991; McLoone, 1994). The range of deprivation scores possible is from -9 to +13, and these are grouped into 7 deprivation categories known as DEPCATs. In this analysis, the DEPCATs have been grouped into 3 blocks - DEPCATs 1 and 2, DEPCATs 3, 4 and 5, and DEPCATs 6 and 7, with 'affluent' defined to be a Carstairs deprivation score less than -3, 'middle' as a score between -3 and 3, and 'deprived' as a score greater than 3 (McLoone and Boddy, 1994).

For the restricted group of approximately 13,600 middle-aged men who attended a third screening visit, data were also available from a full biochemical and haematological profile, as well as measurements of lipids, fibrinogen and plasma viscosity (section 2.4).

Event data obtained from SRL came from the database of Registrar General deaths, hospital discharges (SMR1s) and cancer registrations (SMR6s). SMR6s were only

available until the end of 1992, while deaths and SMR1s were available until the end of 1993. These national data sets have been described in section 3.4.1.

All events which were identified between initial screening (October 1988) and the end of 1993 were included. This resulted in between 3 and 5 years of follow-up on each subject, depending on when the subject was screened. Analysis considered only the subjects who had sufficient identifying information to allow follow-up by record linkage techniques (97,165 out of the total 105,383 subjects).

6.2.2 Categorisation

To maintain subject anonymity, and thus meet the conditions of the linkage approval at SRL, it was necessary to construct categorisations of the data available for the WOSCOPS subjects, on the basis of which morbidity and mortality data would be released by SRL. Each contributing continuous variable was categorised. Each cell in the cross-classification of the categorical variables was required to be of a reasonable size. 'Reasonable' was generally taken to mean that there were at least 5 subjects in each cell. Event records extracted at SRL were then identified only by a cell number, and were not identifiable at an individual subject level.

Many cross-classifications were considered. This thesis will discuss only three, one for each of the subgroups of subjects as given below:

1. A cross-classification of CHD risk factors for middle-aged men at screening visit 1: This cross-classification involved 2 levels of age (45-54 and 55-64), 3 levels of smoking status (never, ex and current), 5 levels of cholesterol (quintiles), 7 levels of alcohol consumption (0, 1-20, 21-30, 31-40, 41-50, 51-60 and >60 units per week), 2 levels of DBP (split at the median value), 2 levels of BMI (split at the median value), and 3 levels of DEPCAT (affluent, middle and deprived). This cross-classification was selected since cholesterol and alcohol were the main factors of interest, and it was thus desirable to incorporate as many levels of them as possible.

2. A cross-classification based on CHD risk factors for women attending screening visit 1: This cross-classification involved 4 levels of age (<45, 45-54, 55-64 and >64), 3 levels of smoking (never, ex and current), 5 levels of cholesterol (quintiles), 3 levels of alcohol (0, 1-10 and >10), 2 levels of DBP (split at the median value), 2 levels of BMI (split at the median value), and 3 levels of DEPCAT (affluent, middle and deprived). This cross-classification was selected to allow an analysis comparable to that already carried out for middle-aged men. Fewer levels of alcohol were possible since only a small number of women consumed large amounts each week.

3. A cross-classification involving laboratory measurements for men who reached screening visit 3: This cross-classification involved 2 levels of age (45-54 and 55-64), 3 levels of smoking (never, ex and current), 2 levels of DBP (split at the median value), 2 levels of LDL cholesterol (split at the median value), 3 levels of HDL cholesterol (tertiles), 3 levels of plasma viscosity (tertiles), and 5 levels of fibrinogen (quintiles). Fibrinogen and viscosity were the main factors of interest in this cross-classification. Smoking was an important factor since it is known to be related to fibrinogen levels.

The need to maintain subject anonymity placed severe restrictions on the number of factors and the number of levels of each factor which could be included in a cross-classification. These cross-classifications were selected by a trial and error process. This process involved experimenting with tabulations in order to identify cross-classifications which allowed the maximum number of factors and levels whilst achieving reasonable cell sizes.

6.2.3 Analysis for each cross-classification

The first stage in each analysis involved tabulations of the number of deaths for various causes of death. The main outcomes of interest were CHD, cancer and injury/poisoning, since some studies have found the incidence of cancer and injury/poisoning to be related to cholesterol level measured at a previous screening visit.

Formal analysis was carried out using logistic regression since the outcome was binary (event / no event), and individual survival times were not available (due to the need to preserve subject anonymity). Alternatives to this method of analysis will be discussed later.

Univariate logistic models were fitted first, to examine the relationship between outcome and each variable on its own. The next stage was to fit logistic regression main effects models, simultaneously including all variables involved in the cross-classification, to assess the contribution of each explanatory variable after the others were included in the model. In addition to p-values for the significance of each variable, 'contrasts' among different levels of the risk factors were studied to investigate in more detail the relationships between the factors and mortality risk. A Bonferroni correction was used within each risk factor to adjust for multiple comparisons. Pairwise contrasts were considered significant if the p-value was less than 0.0167 for factors with 3 levels (such as DEPCAT and smoking), less than 0.005 for factors with 5 levels (such as cholesterol) and less than 0.0024 for factors with 7 levels (such as alcohol).

Finally, interactions were investigated. Interaction terms were added into the logistic model in a forward stepwise manner, in which higher order terms were only allowed once their constituent lower order terms had been included. For example, the age*BMI interaction could only be considered once the age and BMI main effects had been included in the model. Tabulations of the observed relative frequencies of deaths and the expected relative frequencies of deaths based on the logistic main effects model were constructed to investigate the nature of any significant interaction effects. The significance of the parameters fitted in the final models (following stepwise logistic regression), for contrasts between levels of the main effects and interactions, were also examined, to provide further information on the nature of the effects. The coefficient divided by the standard error, for each contrast, is an approximate z-statistic. The exponential of the coefficient provides an estimate of the odds ratio for the levels being contrasted. Plots of residuals against fitted values were constructed to investigate model fit. Large residuals indicate observations which are poorly accounted for by the model (Pregibon, 1981). Chi-square goodness-of-fit statistics and dispersion parameters were also examined to provide some measure of the appropriateness of the fitted model.

6.2.4 Additional analysis for middle-aged men

More detailed analyses were carried out for the group of middle-aged men who were systematically invited to screening visit 1 (see section 2.2), since it had a far larger sample size than any of the other groups considered. The following analyses were carried out.

1) Hospital discharge data were analysed as described in section 6.2.3 for mortality data (that is, univariate logistic regression, multivariate main effects model, forward stepwise logistic regression allowing interactions, tabulations of significant interactions, and model checking). For hospital discharge data there was a sufficient number of events to allow consideration of attempted suicide as a separate analysis outcome. Care was required in the analysis of the SMR1 dataset, since a new SMR1 record is generated every time a subject is hospitalised or moved to another unit (within an episode of stay). The data were analysed on a subject rather than event basis. A subject may be admitted for many different diagnoses. For each diagnosis category, a subject was counted only once in that diagnosis group, no matter how many admissions he had for that diagnosis. A subject could, of course, be included in more than one of the diagnosis groups.

2) Cancer deaths, hospitalisations and registrations were analysed in relation to the cross-classification of CHD risk factors. This was of interest due to the concerns about potential increased cancer risk associated with low cholesterol discussed in section 1.4. Analyses were carried out for 'All malignant neoplasms' and for 'lung neoplasms' - lung being the most common cancer site. For each of these contexts, logistic models were investigated as described in section 6.2.3.

3) Other studies have suggested that the relationship between cholesterol and cancer may be due to undetected disease, which causes the lowering of cholesterol. If this was true, the relationship would diminish with increased length of subject follow-up. To investigate this, the cancer data were analysed as in section 6.2.3, and this was repeated with the first year of follow-up excluded, and then with the first 2 years of follow-up excluded. Although a longer period of follow-up would be required to obtain more

meaningful results (follow-up was only for between 3 and 5 years), this gives an indication of the relationship between cholesterol and cancer once the early years have been excluded. Events which are diagnosed later are less likely to have been pre-existing at baseline.

4) Analysis was also carried out for cancer deaths, with deaths from non-cancer causes excluded, to see whether this strengthened the relationships between the risk factors and outcome. This was only done for total cancer mortality. 1205 subjects died from cancer until the end of 1993. In this analysis, these subjects were considered relative to the 70,909 subjects who had not died from a non-cancer cause. Analysis thus considered factors associated with deaths due to cancer, within the subjects who had not died from other causes.

6.3 Men age 45-64 at screening visit 1

6.3.1 Results

The data for the men age 45-64 were used for the most extensive analyses, since this was the group targeted for possible inclusion in the WOSCOPS clinical trial, and thus made up the largest subgroup screened. There were 74,576 men age 45-64 in the group of 97,165 subjects who were screened and had sufficient identifying information to allow follow-up by record linkage. The mean period of follow-up for the 74,576 middle-aged men was 4.0 years (298,548 subject years in total) until the end of 1993.

6.3.1.1 Preliminary examination of mortality data

The deaths identified by record linkage for the 74,576 men age 45-64 are given in Table 6.1, broken down by cause of death.

Cause of death	Number of deaths
Malignant neoplasms	1230
Coronary heart disease	1333
Cerebrovascular	190
Other circulatory	145
Respiratory	186
Digestive	144
Injury / poisoning	131
Endocrine / nutritional	47
Nervous system	33
Mental disorders	21
Urogenital	10
Infectious disease	5
Blood disorders	5
Other causes	13
Total deaths	**3493**

Table 6.1 Numbers of deaths from various causes for the men age 45-64 at screening visit 1

Table 6.2 contains the frequencies and rates of death for the 74,576 male screenees aged 45-64, broken down by 10-year age band (45-54 and 55-64), smoking status (never, ex or current), cholesterol level (quintiles - 0-4.93, 4.94-5.55, 5.56-6.11, 6.12-6.80 or >6.80 mmol/l), alcohol consumption (0, 1-20, 21-30, 31-40, 41-50, 51-60, or >60 units per week), diastolic blood pressure (< or ≥ the median pressure of 84 mmHg), body mass index (< or ≥ the median of 25.46 kg/m^2) and DEPCAT (affluent, middle or deprived) for all-cause mortality and death due to coronary heart disease, malignant neoplasms and injury/poisoning. The death rates are calculated based on the number of subject years of follow-up accrued within each level of each factor.

Variable	Levels	Number of subjects	Deaths (All-causes)	Deaths (CHD)	Deaths (Cancer)	Deaths (Injury / poisoning)
Age (years)	45-54	37289	887 (5.98)	345 (2.33)	293 (1.97)	52 (0.35)
	55-64	37287	2606 (17.35)	988 (6.58)	937 (6.24)	79 (0.53)
Smoking						
	never	20855	529 (6.38)	212(2.56)	170(2.05)	27(0.33)
	ex-smoker	20773	942(11.29)	377(4.52)	308(3.69)	30(0.36)
	current smoker	32947	2022(15.29)	744(5.63)	752(5.69)	74(0.56)
Cholesterol (mmol/l)	<4.93	14904	838(14.08)	217(3.65)	343(5.76)	37(0.62)
	4.93-5.55	15029	615(10.29)	217(3.63)	233(3.90)	25(0.42)
	5.56-6.11	14890	653(11.00)	253(4.26)	246(4.15)	25(0.42)
	6.12-6.80	14821	618(10.41)	263(4.43)	199(3.35)	13(0.22)
	>6.80	14905	768(12.69)	383(6.33)	208(3.44)	31(0.51)
Alcohol (units/week)	0	15813	929(14.48)	375(5.85)	311(4.85)	32(0.50)
	1-20	43569	1862(10.72)	724(4.17)	675(3.88)	73(0.42)
	21-30	7567	295(9.78)	102(3.38)	118(3.91)	8(0.26)
	31-40	3065	138(11.27)	46(3.76)	46(3.76)	6(0.49)
	41-50	1584	80(12.6)	26(4.10)	26(4.10)	2(0.32)
	51-60	961	60(15.56)	20(5.18)	16(4.15)	6(1.56)
	>60	1270	83(16.35)	23(4.53)	24(4.73)	3(0.59)
DBP (mmHg)	<84	35978	1613 (11.18)	570 (3.95)	620 (4.30)	71 (0.49)
	≥84	38590	1880 (12.17)	763 (4.94)	610 (3.95)	60 (0.39)
BMI (kg/m^2)	<25.46	37236	1946 (13.01)	665 (4.44)	764 (5.11)	81 (0.54)
	≥25.46	37308	1545 (10.37)	667 (4.48)	465 (3.12)	50 (0.34)
DEPCAT						
	affluent	10874	320(7.76)	118(2.86)	125(3.03)	11(0.22)
	middle	42615	1870(10.87)	733(4.26)	665(3.86)	71(0.41)
	deprived	20453	1264(15.27)	472(5.70)	430(5.19)	45(0.54)

Table 6.2: Number of deaths (events/1000 years of follow-up) for middle-aged, male screenees broken down by each of the risk factors. It should be noted that due to a small amount of missing data for each variable the total numbers of subjects are not consistent across the variables.

The results in Table 6.2 parallel those presented in other epidemiological studies, with increased risk of CHD death in smokers (Parish et al, 1995) and in older subjects with raised cholesterol level or blood pressure (see Table 1.1; Smith et al, 1989; Stokes et al, 1987; Martin et al, 1986). Increased risk of death due to cancer or external causes was associated with current smoking and low cholesterol or alcohol consumption in other studies also (Muldoon et al, 1990; Isles et al, 1989; Law et al, 1994). Elevated death rates were also found in other studies to be associated with the group who reported zero alcohol consumption (Friedman et al, 1986), and the group categorised as deprived (Pocock et al, 1987).

6.3.1.2 Logistic models for mortality

Main effects models

The relationship between mortality and the factors listed above was considered for the 73,110 men age 45-64 who had no missing values on the explanatory variables and could thus be used in fitting models. Main effects logistic regression models were fitted for deaths prior to the end of 1993. The odds ratios (of having an event relative to not having an event) and p-values obtained from the univariate (unadjusted) and multivariate (adjusted for all other factors) main effects models are given in table 6.3 for all-cause mortality, table 6.4 for CHD mortality, table 6.5 for cancer mortality and table 6.6 for trauma mortality. It should be noted that the odds ratios are calculated relative to the highest level of each factor and the p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.32		0.32	
	55-64	1.00	<0.0001	1.00	<0.0001
Smoking					
	never	0.40		0.42	
	ex-smoker	0.73		0.70	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	1.10		1.04	
	4.93-5.55	0.79		0.78	
	5.56-6.11	0.85		0.84	
	6.12-6.80	0.81		0.79	
	>6.80	1.00	<0.0001	1.00	<0.0001
Alcohol (units/week)	0	0.89		0.94	
	1-20	0.64		0.69	
	21-30	0.58		0.62	
	31-40	0.67		0.72	
	41-50	0.76		0.77	
	51-60	0.95		1.01	
	>60	1.00	<0.0001	1.00	<0.0001
DBP (mmHg)	<84	0.92		0.89	
	≥84	1.00	0.0156	1.00	0.0012
BMI (kg/m^2)	<25.46	1.28		1.20	
	≥25.46	1.00	<0.0001	1.00	<0.0001
DEPCAT					
	affluent	0.46		0.57	
	middle	0.70		0.76	
	deprived	1.00	<0.0001	1.00	<0.0001

Table 6.3. Odds ratios and p-values for the univariate and multivariate models for all-cause mortality. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of the factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.35		0.35	
	55-64	1.00	<0.0001	1.00	<0.0001
Smoking					
	never	0.45		0.44	
	ex-smoker	0.80		0.73	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	0.57		0.56	
	4.93-5.55	0.56		0.57	
	5.56-6.11	0.66		0.67	
	6.12-6.80	0.70		0.69	
	>6.80	1.00	<0.0001	1.00	<0.0001
Alcohol (units/week)	0	1.35		1.36	
	1-20	0.94		0.97	
	21-30	0.76		0.77	
	31-40	0.85		0.85	
	41-50	0.93		0.89	
	51-60	1.18		1.16	
	>60	1.00	<0.0001	1.00	<0.0001
DBP (mmHg)	<84	0.80		0.81	
	≥84	1.00	0.0001	1.00	0.0003
BMI (kg/m^2)	<25.46	1.01		1.01	
	≥25.46	1.00	0.0916	1.00	0.9225
DEPCAT					
	affluent	0.48		0.54	
	middle	0.76		0.76	
	deprived	1.00	<0.0001	1.00	<0.0001

Table 6.4. Odds ratios and p-values for the univariate and multivariate models for CHD mortality. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.30		0.31	
	55-64	1.00	<0.0001	1.00	<0.0001
Smoking					
	never	0.36		0.38	
	ex-smoker	0.65		0.65	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	1.68		1.53	
	4.93-5.55	1.12		1.06	
	5.56-6.11	1.20		1.13	
	6.12-6.80	0.97		0.95	
	>6.80	1.00	<0.0001	1.00	<0.0001
Alcohol (units/week)	0	1.06		1.08	
	1-20	0.83		0.88	
	21-30	0.83		0.88	
	31-40	0.80		0.86	
	41-50	0.88		0.87	
	51-60	0.89		0.96	
	>60	1.00	0.0300	1.00	0.1165
DBP (mmHg)	<84	1.10		1.00	
	≥84	1.00	0.1845	1.00	0.9645
BMI (kg/m^2)	<25.46	1.61		1.48	
	≥25.46	1.00	<0.0001	1.00	<0.0001
DEPCAT					
	affluent	0.55		0.70	
	middle	0.76		0.84	
	deprived	1.00	<0.0001	1.00	0.0008

Table 6.5. Odds ratios and p-values for the univariate and multivariate models for cancer mortality. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.66		0.61	
	55-64	1.00	0.0081	1.00	0.0070
Smoking	never	0.56		0.66	
	ex-smoker	0.63		0.70	
	current smoker	1.00	0.0239	1.00	0.1070
Cholesterol (mmol/l)	<4.93	1.19		1.09	
	4.93-5.55	0.79		0.79	
	5.56-6.11	0.80		0.84	
	6.12-6.80	0.42		0.45	
	>6.80	1.00	0.0490	1.00	0.0806
Alcohol (units/week)	0	0.84		0.89	
	1-20	0.70		0.72	
	21-30	0.44		0.48	
	31-40	0.82		0.88	
	41-50	0.53		0.55	
	51-60	2.62		2.78	
	>60	1.00	0.0328	1.00	0.0458
DBP (mmHg)	<84	1.24		1.18	
	≥84	1.00	0.1915	1.00	0.3669
BMI (kg/m^2)	<25.46	1.68		1.42	
	≥25.46	1.00	0.0129	1.00	0.0619
DEPCAT	affluent	0.46		0.55	
	middle	0.76		0.85	
	deprived	1.00	0.066	1.00	0.2176

Table 6.6. Odds ratios and p-values for the univariate and multivariate models for trauma mortality. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

To help interpret the data, Bonferroni adjusted 95% confidence intervals for all the pairwise contrasts among the levels of the risk factors were computed. The results are summarised in Table 6.7.

Risk factor	All-cause mortality	Cancer mortality	CHD mortality	Trauma mortality
Age	1 2	1 2	1 2	1 2
Smoking	1 2 3	1 2 3	1 2 3	1 2 3
DBP	1 2	2 1	1 2	2 1
BMI	2 1	2 1	2 1	2 1
Cholesterol	2 4 3 5 1	4 5 2 3 1	1 2 3 4 5	4 2 3 5 1
Alcohol	3 2 4 5 6 7 1	2 3 5 4 6 7 1	2 3 4 5 7 6 1	3 5 2 1 4 7 6
DEPCAT	1 2 3	1 2 3	1 2 3	1 2 3

Table 6.7. Significance of pairwise contrasts among levels of the risk factors in the main effects logistic models for men age 45-64. Levels of each risk factor are given in order of increasing risk from low to high. Levels which are joined by a line were not significantly different from each other after adjusting for multiple comparisons by means of a Bonferroni correction.

Stepwise modelling

The next stage of the analysis was to allow interaction terms to be brought into the logistic model, in a stepwise manner, bringing in the most significant term at each step, given the terms already included. Higher order interactions were considered only when the lower order terms contributing to them had been brought into the model.

For all-cause mortality, this forward stepwise process included all main effects and introduced four interaction terms - smoking*DEPCAT (p=0.003), smoking*BMI (p=0.004), smoking*age (p=0.005) and age*DBP (p=0.024). The parameter estimates obtained by fitting this logistic model are given in Appendix J1, and may be used in

conjunction with the covariance matrix in Appendix F1 to examine any contrast between levels of the factors.

For CHD mortality, the forward stepwise model fitting dropped the BMI main effect and brought in four interaction terms - age*smoking (p=0.004), smoking*DBP (p=0.023), smoking*alcohol (p=0.021) and smoking*DEPCAT (p=0.034).

The parameter estimates, approximate z-statistics (coefficient / standard error) and estimated odds ratios (exponential(coefficient)) obtained by fitting this logistic model are given in Appendix J2. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F2 to examine any contrast among levels of the factors.

For cancer mortality, the stepwise process dropped the main effects for alcohol and DBP. None of the interaction terms were significant although DEPCAT*age (p=0.0639) and alcohol*BMI (p=0.0637) were borderline. For trauma deaths, only the main effects for age and alcohol consumption were included by the stepwise model fitting.

The parameter estimates, approximate z-statistics (coefficient / standard error) and estimated odds ratios (exponential(coefficient)) obtained by fitting the final logistic model are given in Appendix J3 for cancer mortality and Appendix J4 for trauma mortality. These parameter estimates may be used in conjunction with the covariance matrices given in Appendices F3 and F4 to examine any contrast between levels of the factors.

Tabulations of the observed and expected proportions of deaths/subjects in each category were constructed to help investigate the nature of the interaction effects (see Appendix G).

For all-cause mortality, each of the main effect terms was significant. Mortality risk increased with increasing age, blood pressure and deprivation category. Current smokers were at greater risk than ex-smokers, who were, in turn, at greater risk than subjects who had never smoked. Mortality risk was greater with lower levels of BMI. For cholesterol,

the lowest and the highest quintiles were at significantly greater risk than the three middle quintiles. The highest risk alcohol group was consumption of more than 50 units per week, with the second highest risk being for the group consuming 0 units per week. The mortality risk for these groups was only significantly greater than the risk for the groups consuming 1-30 units per week. Significant interaction terms, brought in by the stepwise logistic modelling, were smoking*DEPCAT, smoking*BMI, smoking*age and age*DBP. Mortality risk increased with smoking category (never/ex/current) at each level of DEPCAT, BMI and age (see Appendix G nos 1-4). It increased with DEPCAT and with age at each level of smoking. However, the risk increased with BMI for those who had never smoked but decreased for current and ex-smokers. It increased with DBP for each age group, and increased with age for each DBP category.

CHD mortality was significantly related to each of the main effect terms except BMI. Mortality risk increased with increasing age, blood pressure, cholesterol and deprivation. Risk also increased with smoking status (never/ex/current). The group claiming consumption of 0 units of alcohol per week was at significantly greater risk than the groups consuming 1-30 units per week. Significant interaction terms brought in by the stepwise modelling were age*smoking, smoking*DBP, smoking*alcohol and smoking*DEPCAT. CHD mortality risk increased with smoking category at each level of alcohol consumption, age, DEPCAT and DBP. It generally decreased with alcohol and increased with age, DBP and DEPCAT for each smoking category (see Appendix G nos 5-8).

Cancer mortality was not significantly related to DBP or alcohol consumption. However, cancer risk did increase with increasing age, deprivation and smoking status (never/ex/current). Cancer risk was also greater for lower BMI, and for the lowest quintile of cholesterol. No interaction terms were significant for cancer mortality.

Trauma mortality was significantly related to only age and alcohol consumption, with increased risk for the older age group, and the higher levels of alcohol consumption, although none of the pairwise contrasts between levels of alcohol consumption were significant. No interaction terms were significant for trauma mortality.

As a further means of investigating the interaction effects, plots were constructed for the fitted values at relevant cross-tabulations of factors. This was done using the final stepwise models with factors not involved in the interaction of interest kept at baseline levels. This gives a pictorial representation of the nature of the interaction effect.

For all-cause mortality, the smoking*DEPCAT interaction is illustrated in Figure 6.1, the smoking*BMI interaction in Figure 6.2, the smoking*age interaction in Figure 6.3 and the age*DBP interaction in Figure 6.4.

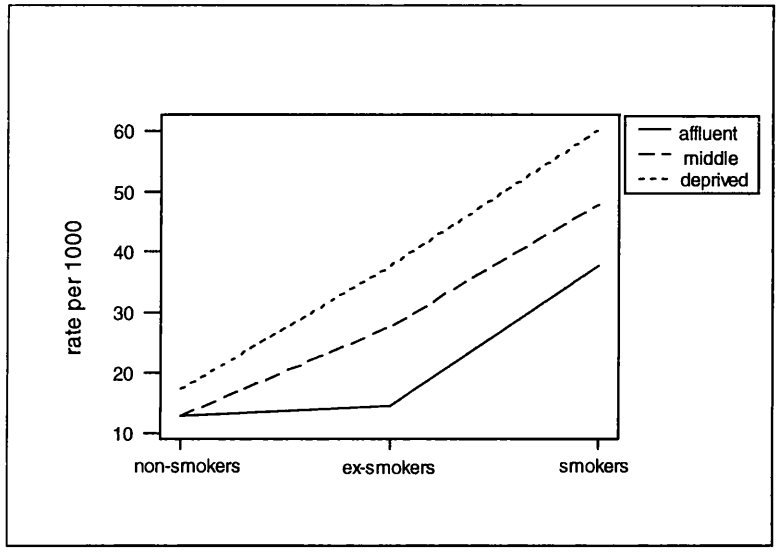


Figure 6.1. Fitted values from the final stepwise model for all-cause mortality at cross-tabulations of smoking*DEPCAT.

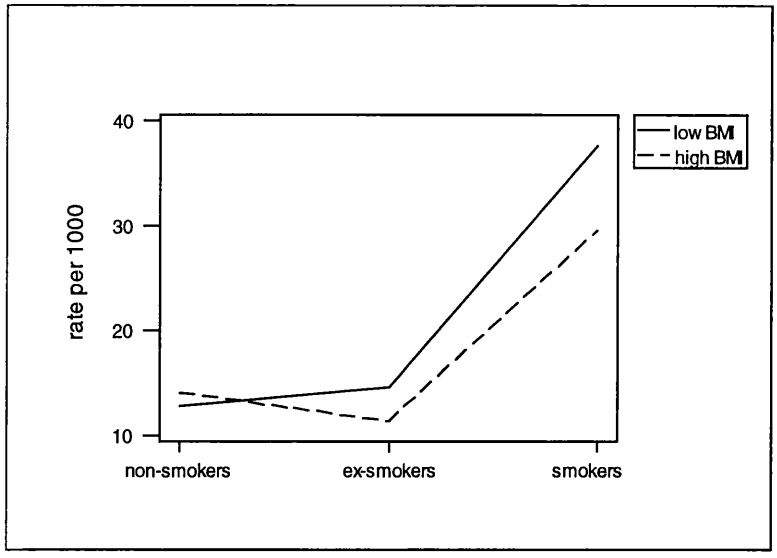


Figure 6.2. Fitted values from the final stepwise model for all-cause mortality at cross-tabulations of smoking*BMI.

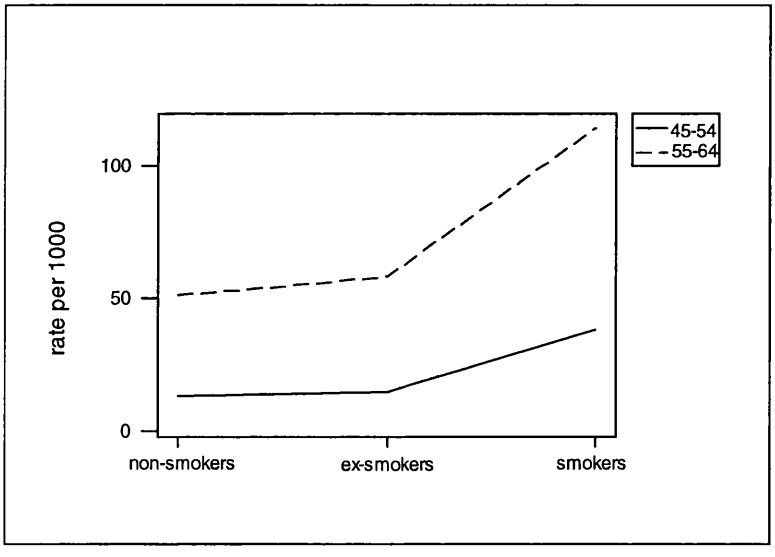


Figure 6.3. Fitted values from the final stepwise model for all-cause mortality at cross-tabulations of smoking*age.

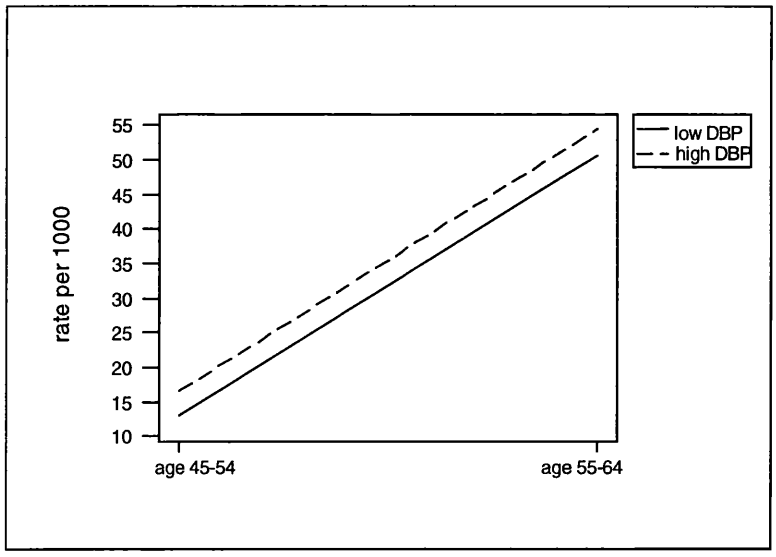


Figure 6.4. Fitted values from the final stepwise model for all-cause mortality at cross-tabulations of age*DBP.

The significant interaction between DEPCAT and smoking for all-cause mortality may relate to lower smoking rates among affluent compared to deprived subjects. There is higher mortality in middle and deprived ex-smokers than would have been expected (Figure 6.1). The age*smoking interaction is due to a wider diference in mortality for smokers compared to non and ex-smokers (Figure 6.3). The age*DBP interaction seems to be picking up a very small effect, with the graph showing little of real interest (Figure 6.4). Mortality risk increased with BMI for the subgroup of men who had never smoked, but decreased with BMI for current and ex-smokers (Figure 6.2). This interaction may be due to current smokers being at greater risk of lung cancer which is also related to low BMI. Thus, obesity appears to be a risk factor in the absence of smoking, but smoking is a more important factor and is related to low BMI.

For CHD mortality, the smoking*age interaction is illustrated in Figure 6.5, the smoking*DBP interaction in Figure 6.6, the smoking*alcohol interaction in Figure 6.7 and the smoking*DEPCAT interaction in Figure 6.8.

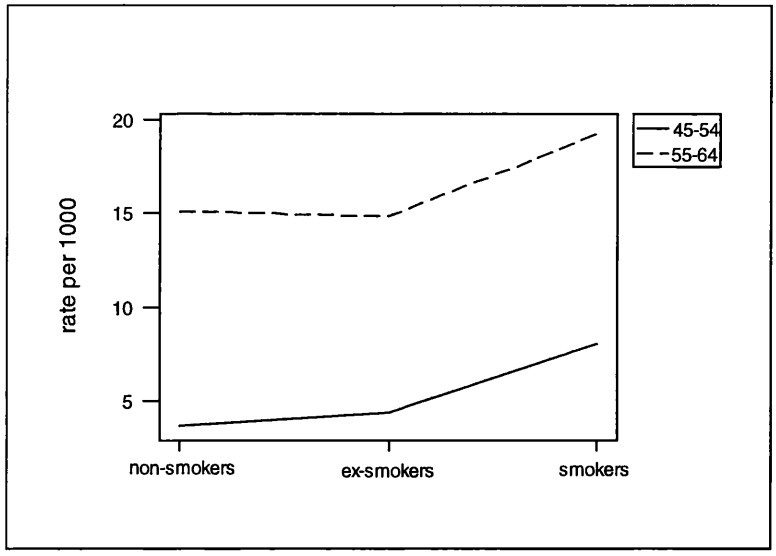


Figure 6.5. Fitted values from the final stepwise model for CHD mortality at cross-tabulations of smoking*age.

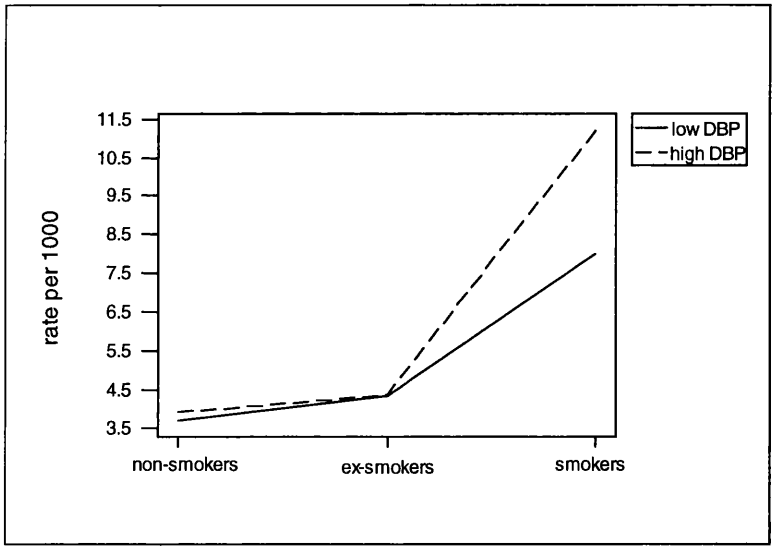


Figure 6.6. Fitted values from the final stepwise model for CHD mortality at cross-tabulations of smoking*DBP.

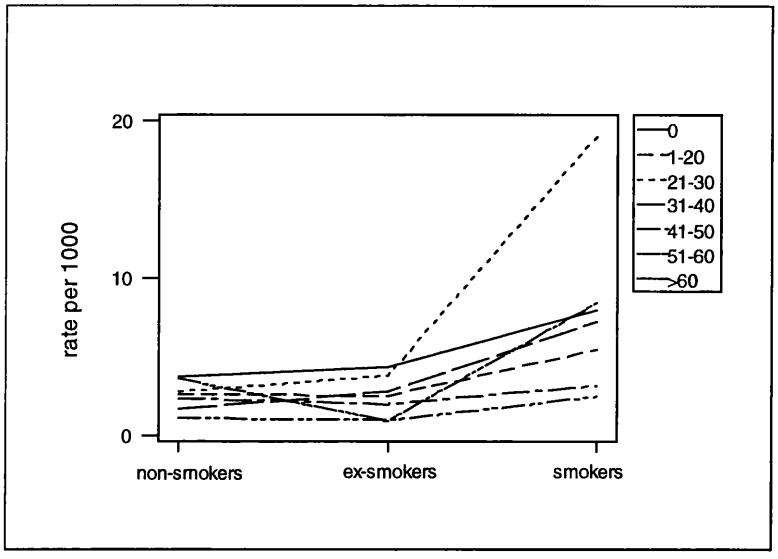


Figure 6.7. Fitted values from the final stepwise model for CHD mortality at cross-tabulations of smoking*alcohol.

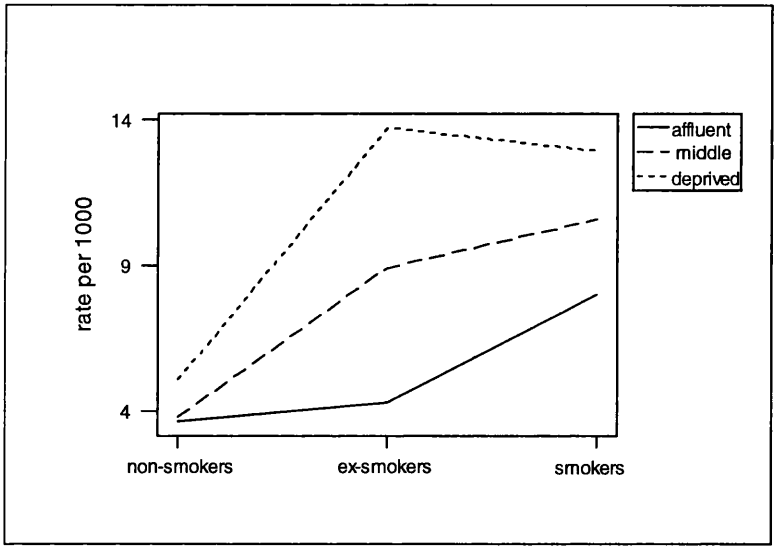


Figure 6.8. Fitted values from the final stepwise model for CHD mortality at cross-tabulations of smoking*DEPCAT.

The difference between ex-smokers and men who have never smoked was approximately the same in each of the age groups (see Appendix G no 5). However, the CHD mortality rate observed for the current smokers was around three times that in the never smokers for the younger age group but only around twice that observed for never smokers for the older age group. The graph shows similar trends (Figure 6.5). For the DBP*smoking interaction, there was little DBP difference for never and ex-smokers, while for current smokers there was greater mortality risk for subjects with high DBP. For alcohol and smoking, current smokers consuming 21-30 units of alcohol per week were at substantially greater risk of CHD mortality than subjects at any other level of these risk factors. The difference between the levels of DEPCAT are similar for current smokers and men who have never smoked. However, the differences are much larger for the group of ex-smokers. The proportion of CHD deaths for ex-smokers in the middle DEPCAT is approximately twice that for affluent ex-smokers, while the proportion for deprived ex-smokers is around three times that for affluent ex-smokers (see Appendix G no 8).

The fit of the final models was checked by means of plots of residuals against fitted values, which were expected to show a random scatter around 0. As an example of the model fit achieved, the residual plot for the outcome of all-cause mortality is given in Appendix H. Chi-square goodness-of-fit tests leading to estimates of the dispersion parameters were also examined (see Appendix H) but did not indicate that there were any problems with model fit.

6.3.1.3 Logistic models for hospitalisation

An approach to the analysis of morbidity was to use the SMR1 (hospital discharge) database which went up to the end of 1993. The numbers of 45-64 year old men identified by record linkage as being hospitalised for various events are given in Table 6.8. It should be noted that the diagnosis of 'mental disorders' will usually be recorded on the SMR4 (psychiatric hospital admission) form rather then an SMR1 form.

Main effects models

As for mortality, to give an overview, logistic main effects models were fitted for the 73,110 men age 45-64 who had sufficient identifying information to allow follow-up by record linkage and no missing values for any of the explanatory variables contributing to the categorisation. The odds ratios (of having an event relative to not having an event) and p-values obtained from the univariate (unadjusted) and multivariate (adjusted for all other factors) main effects models are given in table 6.9 for all-cause hospitalisation, table 6.10 for CHD hospitalisation, table 6.11 for cancer hospitalisation, table 6.12 for trauma hospitalisation and table 6.13 for suicide hospitalisation. It should be noted that the odds ratios are calculated relative to the highest level of each factor and the p-values relate to the overall significance of each factor.

Reason for hospitalisation	Number of SMR1 records	Number of subjects
Malignant neoplasms	8765	2626
Other neoplasms	1647	1209
Coronary heart disease	9722	4037
Cerebrovascular	1298	906
Other circulatory	6439	3783
Respiratory	2967	1984
Digestive	9007	6010
Injury / poisoning	3571	2103
Endocrine / nutritional	1130	475
Nervous system	2992	2118
Mental disorders	410	298
Urogenital	5331	3530
Infectious disease	174	132
Blood disorders	602	303
Congenital	134	115
Musculoskeletal	3432	2490
Skin disease	2234	1722
Miscellaneous	10994	7534
Total	**70849**	**26693**

Table 6.8 Numbers of hospitalisations from various causes for the men age 45-64. The miscellaneous category related to a primary diagnosis of a 'symptom' or a 'supplementary' code of factors influencing contact with health services and thus did not represent events which were meaningful in themselves. It should be noted that the total number of subjects is not equal to the sum of the column since subjects may be included in more than one diagnosis group.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.95		0.63	
	55-64	1.00	<0.0001	1.00	<0.0001
Smoking	never	0.72		0.72	
	ex-smoker	0.94		0.92	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	0.98		0.98	
	4.93-5.55	0.88		0.88	
	5.56-6.11	0.90		0.90	
	6.12-6.80	0.91		0.91	
	>6.80	1.00	<0.0001	1.00	<0.0001
Alcohol (units/week)	0	1.04		1.04	
	1-20	0.80		0.82	
	21-30	0.77		0.79	
	31-40	0.81		0.82	
	41-50	0.75		0.74	
	51-60	0.92		0.94	
	>60	1.00	<0.0001	1.00	<0.0001
DBP (mmHg)	<84	1.00		1.00	
	≥84	1.00	0.9048	1.00	0.8991
BMI (kg/m^2)	<25.46	1.01		0.99	
	≥25.46	1.00	0.4599	1.00	0.6047
DEPCAT	affluent	0.67		0.73	
	middle	0.84		0.87	
	deprived	1.00	<0.0001	1.00	<0.0001

Table 6.9. Odds ratios and p-values for the univariate and multivariate models for all-cause hospitalisation. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.57		0.60	
	55-64	1.00	<0.0001	1.00	<0.0001
Smoking					
	never	0.55		0.53	
	ex-smoker	0.99		0.88	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	0.39		0.41	
	4.93-5.55	0.53		0.55	
	5.56-6.11	0.68		0.70	
	6.12-6.80	0.75		0.76	
	>6.80	1.00	<0.0001	1.00	<0.0001
Alcohol (units/week)	0	2.00		2.03	
	1-20	1.44		1.45	
	21-30	1.10		1.09	
	31-40	1.28		1.25	
	41-50	1.00		0.95	
	51-60	1.13		1.10	
	>60	1.00	<0.0001	1.00	<0.0001
DBP (mmHg)	<84	0.86		0.90	
	≥84	1.00	<0.0001	1.00	0.0022
BMI (kg/m^2)	<25.46	0.74		0.76	
	≥25.46	1.00	<0.0001	1.00	<0.0001
DEPCAT					
	affluent	0.63		0.68	
	middle	0.88		0.87	
	deprived	1.00	<0.0001	1.00	<0.0001

Table 6.10. Odds ratios and p-values for the univariate and multivariate models for CHD hospitalisation. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	0.32		0.32	
	55-64	1.00	<0.0001	1.00	<0.0001
Smoking					
	never	0.51		0.52	
	ex-smoker	0.79		0.76	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	1.33		1.28	
	4.93-5.55	1.03		1.01	
	5.56-6.11	1.00		0.99	
	6.12-6.80	0.94		0.93	
	>6.80	1.00	<0.0001	1.00	<0.0001
Alcohol (units/week)	0	0.92		0.86	
	1-20	0.78		0.76	
	21-30	0.74		0.74	
	31-40	0.81		0.83	
	41-50	0.68		0.65	
	51-60	0.91		0.94	
	>60	1.00	0.0051	1.00	0.0351
DBP (mmHg)	<84	1.04		1.01	
	≥84	1.00	0.3672	1.00	0.7697
BMI (kg/m^2)	<25.46	1.31		1.21	
	≥25.46	1.00	<0.0001	1.00	<0.0001
DEPCAT					
	affluent	0.73		0.89	
	middle	0.77		0.85	
	deprived	1.00	<0.0001	1.00	0.0010

Table 6.11. Odds ratios and p-values for the univariate and multivariate models for cancer hospitalisation. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	1.13		1.08	
	55-64	1.00	0.0072	1.00	0.0744
Smoking					
	never	0.65		0.76	
	ex-smoker	0.62		0.71	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	1.36		1.25	
	4.93-5.55	1.06		1.02	
	5.56-6.11	1.02		1.00	
	6.12-6.80	0.98		0.98	
	>6.80	1.00	<0.0001	1.00	0.0007
Alcohol (units/week)	0	0.32		0.38	
	1-20	0.33		0.39	
	21-30	0.42		0.48	
	31-40	0.50		0.54	
	41-50	0.54		0.58	
	51-60	0.73		0.78	
	>60	1.00	<0.0001	1.00	<0.0001
DBP (mmHg)	<84	1.02		0.98	
	≥84	1.00	0.6113	1.00	0.5823
BMI (kg/m^2)	<25.46	1.40		1.31	
	≥25.46	1.00	<0.0001	1.00	<0.0001
DEPCAT					
	affluent	0.44		0.50	
	middle	0.70		0.76	
	deprived	1.00	<0.0001	1.00	<0.0001

Table 6.12. Odds ratios and p-values for the univariate and multivariate models for trauma hospitalisation. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

Variable	Levels	Univariate		Multivariate	
		Odds Ratio	p-value	Odds ratio	p-value
Age (years)	45-54	1.46		1.37	
	55-64	1.00	0.0112	1.00	0.0357
Smoking					
	never	0.29		0.36	
	ex-smoker	0.39		0.49	
	current smoker	1.00	<0.0001	1.00	<0.0001
Cholesterol (mmol/l)	<4.93	1.34		1.13	
	4.93-5.55	1.01		0.94	
	5.56-6.11	1.06		1.03	
	6.12-6.80	1.16		1.15	
	>6.80	1.00	0.6476	1.00	0.8982
Alcohol (units/week)	0	0.30		0.42	
	1-20	0.16		0.22	
	21-30	0.42		0.52	
	31-40	0.41		0.47	
	41-50	0.30		0.34	
	51-60	0.30		0.34	
	>60	1.00	<0.0001	1.00	<0.0001
DBP (mmHg)	<84	1.38		1.26	
	≥84	1.00	0.0294	1.00	0.1223
BMI (kg/m^2)	<25.46	1.79		1.50	
	≥25.46	1.00	0.0001	1.00	0.0098
DEPCAT					
	affluent	0.29		0.39	
	middle	0.54		0.64	
	deprived	1.00	<0.0001	1.00	0.007

Table 6.13. Odds ratios and p-values for the univariate and multivariate models for suicide hospitalisation. All odds ratios are calculated relative to the highest level of each factor and p-values relate to the overall significance of each factor.

To see which levels of the risk factors were related to increased risk of a subject being hospitalised, the pairwise contrasts among the levels of each risk factor were computed. A Bonferroni correction was applied within each risk factor to adjust for multiple comparisons with results as in Table 6.14.

Risk factor	All-cause	CHD	Cancer	Trauma	Suicide
Age	1 2	1 2	1 2	2 1	2 1
Smoking	1 2 3	1 2 3	1 2 3	2 1 3	1 2 3
DBP	2 1	1 2	2 1	1 2	2 1
BMI	1 2	1 2	2 1	2 1	2 1
Cholesterol	2 3 4 1 5	1 2 3 4 5	4 3 5 2 1	4 3 5 2 1	2 5 3 1 4
Alcohol	5 3 2 4 6 7 1	5 3 7 6 4 2 1	5 3 2 4 6 7 1	1 2 3 5 4 6 7	2 6 5 1 4 3 7
DEPCAT	1 2 3	1 2 3	2 1 3	1 2 3	1 2 3

Table 6.14 Significance of pairwise contrasts among levels of the risk factors for multivariate main effects logistic models for subjects hospitalised. Levels of each risk factor are given in order of increasing risk from low to high. Levels which are joined by a line were not significantly different from each other, after adjusting for multiple comparisons by means of a Bonferroni correction.

Stepwise modelling

The next stage of the analysis was to allow interaction terms to be brought into the logistic model, in a stepwise manner, bringing in the most significant term at each step given the terms already included. Higher order interactions were considered only when the lower order terms contributing to them had all been brought into the model.

For all-cause hospitalisation, this forward stepwise process dropped the DBP and BMI main effects and introduced two interaction terms - DEPCAT*age (p=0.0002) and smoking*age (p=0.026).

The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for hospitalisation from any cause are given in Appendix J5 and may be used in conjunction with the covariance matrix given in Appendix F5 to examine any contrast between levels of the factors.

For CHD hospitalisation, the final model resulting from the forward stepwise model fitting included all the main effects and four interaction terms - age*smoking ($p<0.001$), age*DBP ($p=0.008$), age*cholesterol ($p=0.016$) and age*DEPCAT ($p=0.049$).

The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for CHD hospitalisation are given in Appendix J6 and may be used in conjunction with the covariance matrix given in Appendix F6 to examine any contrast between levels of the factors.

For cancer hospitalisation, none of the interaction terms were significant and the final model contained only the main effects of age, smoking, BMI, cholesterol, DEPCAT and alcohol. The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for cancer hospitalisation are given in Appendix J7 and may be used in conjunction with the covariance matrix given in Appendix F7 to examine any contrast between levels of the factors.

For hospitalisations for injury or poisoning, the stepwise process dropped the terms for the age and DBP main effects and brought in two interaction terms - BMI*DEPCAT ($p=0.007$) and BMI*smoking ($p=0.028$).

The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for trauma hospitalisation are given in Appendix J8 and may be used in conjunction with the

covariance matrix given in Appendix F8 to examine any contrast between levels of the factors.

For hospitalisations for attempted suicide, the final stepwise model dropped the main effects of cholesterol and DBP and introduced the smoking*DEPCAT interaction (p=0.039). The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for hospitalisation for attempted suicide are given in Appendix J9 and may be used in conjunction with the covariance matrix given in Appendix F9 to examine any contrast between levels of the factors.

Tabulations of the observed and expected proportions of deaths/subjects in each category were carried out to investigate the nature of the interaction effects.

Risk of hospitalisation for any cause was not significantly related to DBP or BMI. Risk increased with increasing age, deprivation and smoking status (never/ex/current). The highest risk for cholesterol was for the lowest and highest quintiles, and the greatest risk associated with alcohol was for consumption of 0 or more than 60 units per week. Significant interaction terms brought in by the stepwise modelling were age*DEPCAT and age*smoking. Risk increased with age for each DEPCAT and smoking category, and increased with DEPCAT and smoking category for each age group (see Appendix G nos 9 and 10).

Hospitalisation for CHD was significantly related to each of the main effect terms fitted. Risk increased with increasing age, blood pressure, BMI, cholesterol, deprivation and smoking status (never/ex/current). The group who claimed to have an alcohol consumption of 0 units per week were at significantly greater risk than any of the other alcohol groups. The interaction terms brought in by the stepwise modelling were for age*smoking, age*DBP, age*cholesterol and age*DEPCAT. CHD risk increased with age for each category of smoking, DEPCAT, DBP and cholesterol. It also increased with DBP, cholesterol, DEPCAT and smoking category for each age group (see Appendix G nos 11-14).

Cancer hospitalisation was not significantly related to DBP. However, cancer risk did significantly increase with increasing age and smoking status. Risk was greater for lower levels of BMI, and for the lowest quintile of cholesterol. The most deprived group was at significantly greater risk than the 'middle' group, and the highest alcohol risk was with consumption of 0 units per week, although none of the pairwise contrasts between levels of alcohol consumption were significant. No interaction terms were significant for cancer hospitalisation.

Trauma (injury or poisoning) risk was not significantly related to age or blood pressure. Risk increased with increasing deprivation and alcohol consumption. There was increased risk associated with current smokers, low BMI, and the lowest quintile of cholesterol. Stepwise modelling brought in the interaction terms BMI*DEPCAT and BMI*smoking. Risk decreased with BMI in each DEPCAT and smoking category and increased with DEPCAT and smoking category for each level of BMI (see Appendix G nos 15 and 16).

Hospitalisation for attempted suicide was not significantly related to DBP or cholesterol. Risk was greater for the younger age group, the current smokers, the group with lower BMI and the most deprived group. There was no clear pattern in the risk associated with alcohol consumption, although the lowest risk group was for a consumption of 1-20 units per week. There was also a significant interaction between smoking and DEPCAT. Suicide risk increased with smoking category in the middle and deprived categories of DEPCAT while there was no clear trend in the affluent group. In a similar way, risk increased with DEPCAT for current and ex-smokers but did not show this trend for the group of subjects who had never smoked (see Appendix G no 17).

As a further means of investigating the interaction effects, plots were constructed for the fitted values at relevant cross-tabulations of factors. This was done using the final stepwise models with factors not involved in the interaction of interest kept at baseline levels. This gives a pictorial representation of the nature of the interaction effect.

For all-cause hospitalisation, the DEPCAT*age interaction is illustrated in Figure 6.9 and the smoking*age interaction in Figure 6.10.

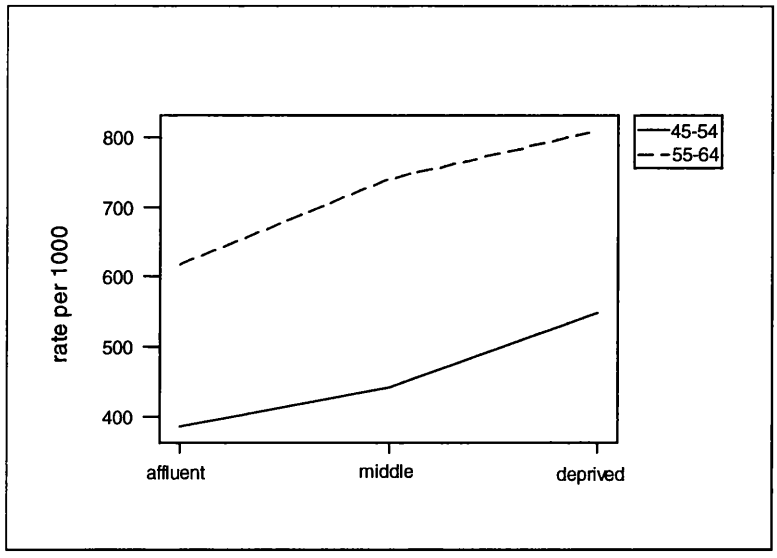


Figure 6.9. Fitted values from the final stepwise model for all-cause hospitalisation at cross-tabulations of DEPCAT*age.

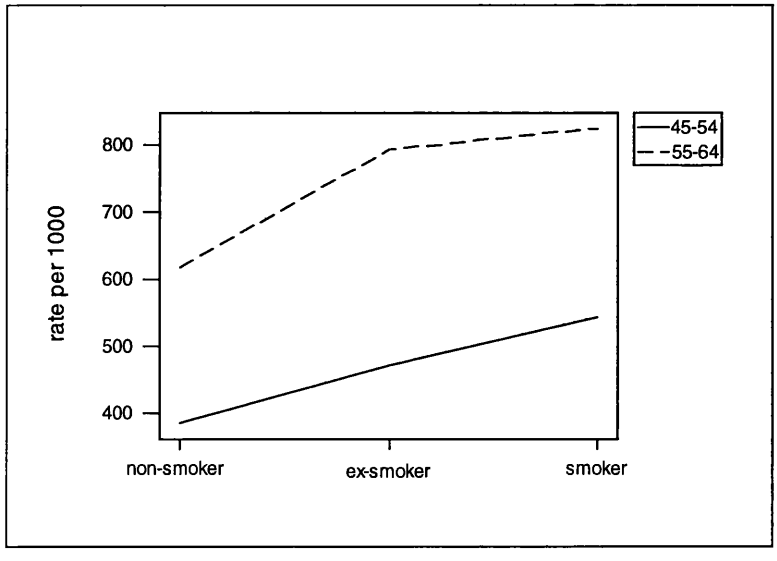


Figure 6.10. Fitted values from the final stepwise model for all-cause hospitalisation at cross-tabulations of smoking*age.

The trends across DEPCAT for each age group are similar (Figure 6.9) and the age*DEPCAT interaction seems to be picking up a very small effect. Similarly, the smoking*age interaction seems to be a marginal effect, although there is higher than expected risk in the older ex-smokers - possibly due to having accumulated more years of smoking before they stopped (Figure 6.10).

For CHD hospitalisation, the age*smoking interaction is illustrated in Figure 6.11, the age*DBP interaction in Figure 6.12, that age*cholesterol interaction in Figure 6.13 and the age*DEPCAT interaction in Figure 6.14.

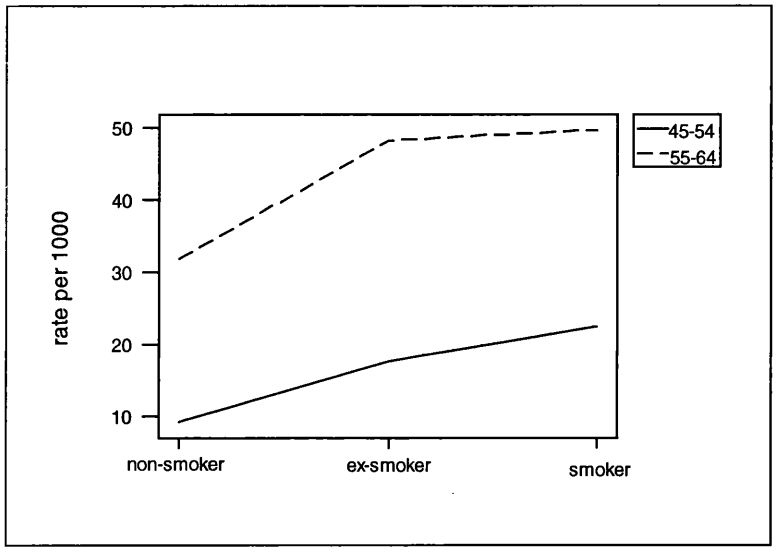


Figure 6.11. Fitted values from the final stepwise model for CHD hospitalisation at cross-tabulations of age*smoking.

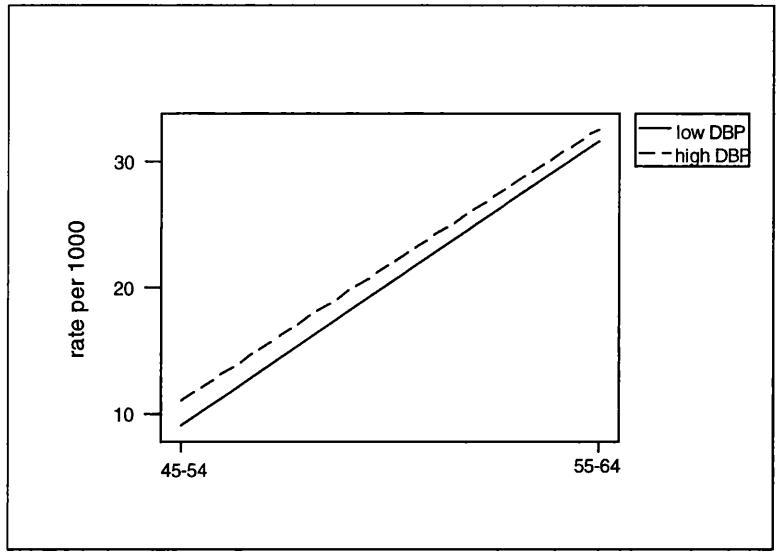


Figure 6.12. Fitted values from the final stepwise model for CHD hospitalisation at cross-tabulations of age*DBP.

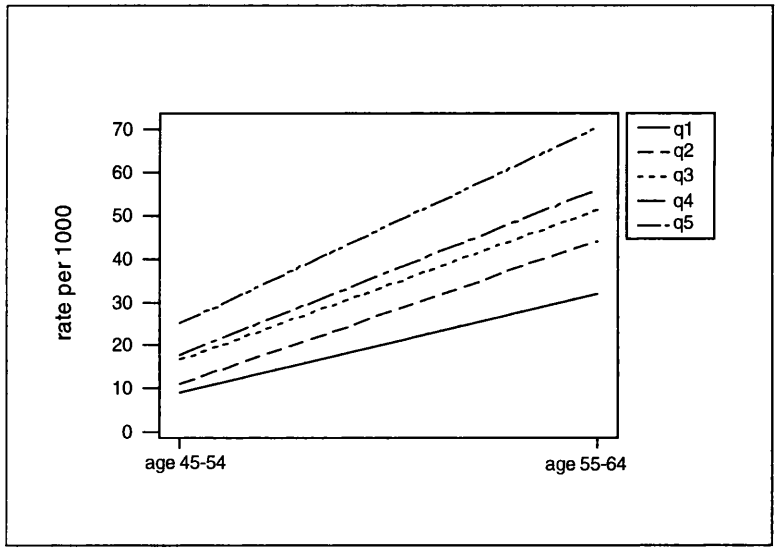


Figure 6.13. Fitted values from the final stepwise model for CHD hospitalisation at cross-tabulations of age*cholesterol.

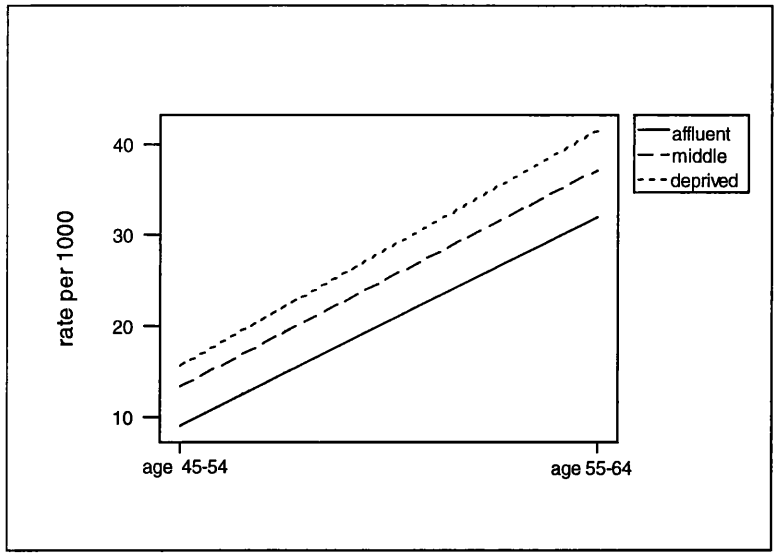


Figure 6.14. Fitted values from the final stepwise model for CHD hospitalisation at cross-tabulations of age*DEPCAT.

For the age*smoking interaction, the pattern was the same as for all-cause hospitalisation (Figure 6.11). The differences between ex-smokers and never smokers and between current smokers and never smokers were larger in the 45-54 year old age group (2 and 2.5 times as big) than in the 55-64 year old group (1.5 and 1.4 times as big - see Appendix G no 11). This is in agreement with a recently published case-control study (Parish et al, 1995) which found that at ages 30-49 the rate of myocardial infarction in smokers was about five times that in non-smokers, while at ages 50-59 the rate was only around three times as big, and at ages 60-79 it was only around twice as large as that in non-smokers. For age*DBP, the difference between the blood pressure groups is slightly larger in the younger age group (Figure 6.12). For the interaction between age and cholesterol, the differences in mortality risk are wider for the older age group than the younger (Figure 6.13). The final interaction for CHD hospitalisation related to age and DEPCAT. This seems to be picking up a small effect with the trend lines being reasonably parallel (Figure 6.14).

For trauma hospitalisation, the DEPCAT*BMI interaction is illustrated in Figure 6.15 and the smoking*BMI interaction in Figure 6.16.

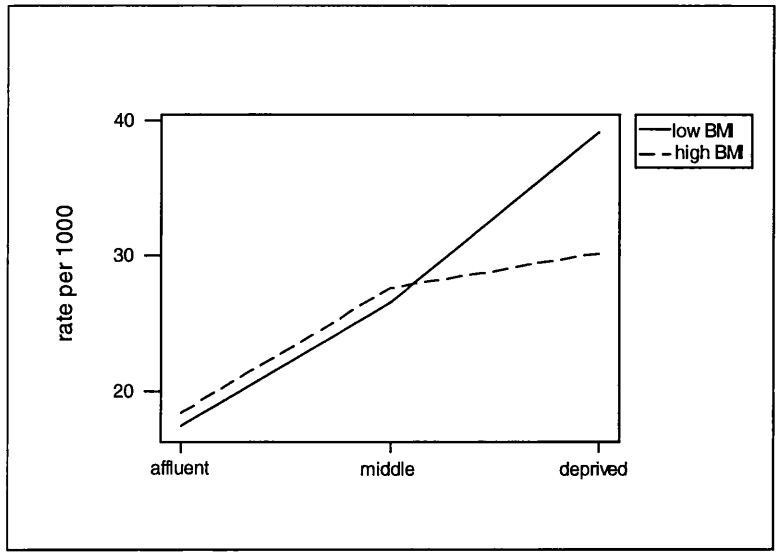


Figure 6.15. Fitted values from the final stepwise model for trauma hospitalisation at cross-tabulations of DEPCAT*BMI.

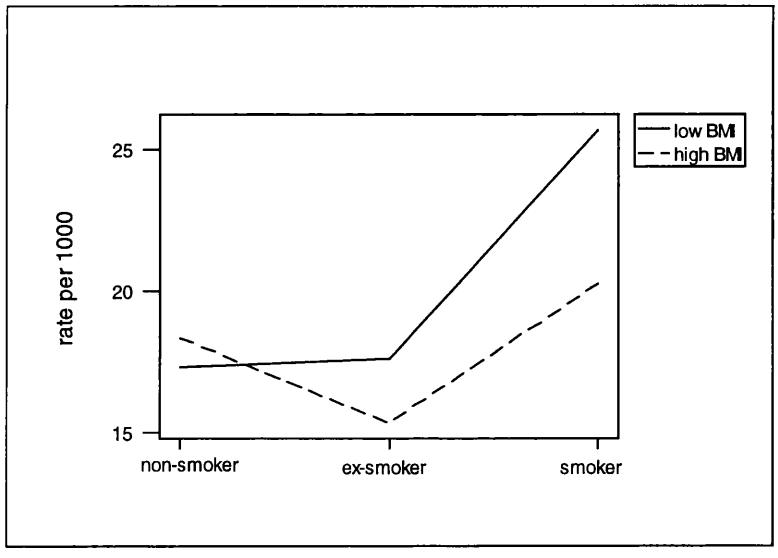


Figure 6.16. Fitted values from the final stepwise model for trauma hospitalisation at cross-tabulations of smoking*BMI.

The proportion of subjects hospitalised for injury or poisoning decreased with BMI but increased with DEPCAT and smoking, each of which had significant interactions with BMI. So, thin men who smoked, and came from the deprived group, were at the highest

risk of hospitalisation for injury or poisoning. For the BMI*DEPCAT interaction, there was little difference in BMI risk for the affluent and middle groups, while for deprived subjects there was much greater risk with low BMI (Figure 6.15). Greater risk was related to high BMI for non-smokers, but low BMI for ex and current smokers (Figure 6.16).

For suicide hospitalisation, the smoking*DEPCAT interaction is illustrated in Figure 6.17. There was a much greater risk difference among the deprivation levels for current smokers than for never and ex-smokers, with the middle deprivation group being at highest suicide risk for smokers, and the deprived group being at the highest risk for never and ex-smokers.

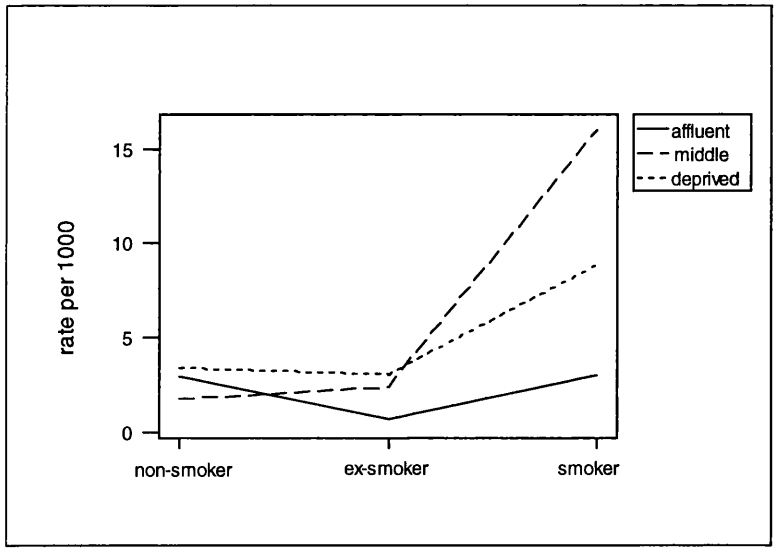


Figure 6.17. Fitted values from the final stepwise model for suicide hospitalisation at cross-tabulations of smoking*DEPCAT.

The fit of the final models was examined by the construction of residual plots and consideration of estimates of the dispersion parameters (see Appendix H). There were no indications that there was a problem with model fit.

6.3.1.4 Comparison of results from mortality, registration and hospitalisation databases

For cancers, there was an alternative database at SRL which provided information on incidence - the Cancer Registration (SMR6) database, which may potentially provide better incidence information than consideration of hospital discharges, since each cancer will be registered only once, at the time of diagnosis. However, cancer registrations only went up to the end of 1992 on the linked database at SRL. Based on cancer registrations until the end of 1992, there were 1510 registrations for the middle-aged men. Of these 1510 cancers, the most common cancer sites were as in Table 6.15.

Cancer Site	Frequency
Lung	362
Skin	204
Bladder	99
Colon	93
Rectum	78
Prostate	76
Stomach	67
Oesophagus	56

Table 6.15 Frequency of cancer sites for men age 45-64 from SMR6s until the end of 1992

Since skin neoplasms (with the exception of malignant melanomas) are often less serious, and are known to be inconsistently registered, lung was the only site with sufficient numbers to be worth consideration on its own at this stage. Further consideration was given to cancers, comparing the results for mortality and incidence, where incidence data were derived from both the SMR1 and SMR6 databases. The number of events accumulated on each record type are given in Table 6.16. Results may differ between the SMR6 and SMR1 analyses since hospitalisation will be required for

the more serious cancers, with 'minor' skin neoplasms being removed by GPs and not appearing on an SMR1. It is also possible that the SMR1 database will contain further treatment admissions for cancers which were diagnosed prior to subject screening, and should thus not be included in prospective follow-up, but these events cannot be distinguished from the information available.

Men age 45-64	73,110
Cancer deaths until the end of 1993	1205
Cancer registrations until the end of 1992	1510
Subjects hospitalised for cancer until the end of 1993	2626

Table 6.16 Numbers of cancer events for men age 45-64

Logistic main effects models were fitted for the categorised records for the diagnosis of all malignant neoplasms for each of the record types and gave the results in Table 6.17.

Risk factor	Deaths	SMR6s	SMR1s
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	<0.0001	<0.0001
DBP	0.9645	0.4759	0.7697
BMI	<0.0001	<0.0001	<0.0001
Cholesterol	<0.0001	0.0605	<0.0001
Alcohol	0.1165	0.0439	0.0351
DEPCAT	0.0008	0.0001	0.0010

Table 6.17. P-values from logistic main effects models for all cancers recorded on each of the databases. P-values relate to the significance of the factor once all other terms have been included.

Stepwise logistic regression for each of the three record types resulted in the models given in Table 6.18.

Explanatory	Deaths	SMR6s	SMR1s
Smoking	<0.0001	<0.0001	<0.0001
Age	<0.0001	<0.0001	<0.0001
BMI	<0.0001	<0.0001	<0.0001
DEPCAT	<0.0001	0.0001	0.0011
Alcohol	-	0.042	0.0376
Cholesterol	<0.0001	0.068	<0.0001
Alcohol*Cholesterol	-	0.0168	-

Table 6.18 P-values for models arising out of stepwise logistic regression for all malignant neoplasms. The p-values are for each effect once all other terms have already been included.

While the results from analysis of the three databases are similar, there are some differences, which will be discussed in section 6.3.2.

The above analysis for each database was repeated for the restricted outcome of lung cancer, for which there were 461 deaths until the end of 1993, 362 registrations until the end of 1992, and 633 subjects hospitalised until the end of 1993. Lung cancer thus accounted for approximately $\frac{1}{3}$ of the total cancer events. Logistic main effects models including all factors were fitted, and resulted in the p-values given in Table 6.19.

Risk Factor	Deaths	SMR6s	SMR1s
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	<0.0001	<0.0001
DBP	0.6490	0.9901	0.7583
BMI	<0.0001	<0.0001	<0.0001
Cholesterol	0.0833	0.0078	0.0356
Alcohol	0.4643	0.7391	0.6687
DEPCAT	0.0073	0.0146	0.0014

Table 6.19 P-values from logistic main effects models for lung cancer. The p-values are for each factor once all other terms have been included.

When stepwise logistic regression was carried out, no interaction terms were brought into the model. Lung cancer gave results which were similar to those for all cancers, although some differences appeared for cholesterol, alcohol and DEPCAT. For total cancers, cholesterol was significant for deaths and SMR1s but not for SMR6s. In contrast, for lung cancer, cholesterol was not significant for deaths, but was for SMR1s and SMR6s. Alcohol was not a significant risk factor for lung cancer whereas it had been significant for total cancer incidence. DEPCAT remained significant for each database, with the affluent group being at the lowest risk, and the deprived group at the highest.

6.3.1.5 Exclusion of early years of follow-up

In order to investigate the hypothesis that low cholesterol is caused by pre-clinical cancer and the relationship between low cholesterol and cancer would thus disappear after a few years of follow-up, the first two years of subject follow-up were excluded, to see whether the strength of the relationship diminished. Only the first two years were excluded since WOSCOPS only had between 3 and 5 years of follow-up until the end of 1993. Total cancer mortality was the outcome considered in this analysis, since specific cancer sites would have had insufficient numbers of events once the early years were excluded.

Logistic main effects models including all factors were fitted for:

1) the 1205 cancer deaths occurring in the 73,110 men age 45-64 (as considered earlier)

2) the 1033 deaths occurring after the first year for the 72,938 men age 45-64 who were followed up for more than one year

3) the 752 deaths occurring after two years of follow-up for the 72,657 men age 45-64 who were followed up for more than two years.

The resulting p-values were as in Table 6.20.

Risk factor	Full follow-up	Year 1 excluded	Years 1 & 2 excluded
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	<0.0001	<0.0001
DBP	0.9645	0.6131	0.9816
BMI	<0.0001	<0.0001	<0.0001
Cholesterol	<0.0001	0.0008	0.0693
Alcohol	0.1165	0.2013	0.8712
DEPCAT	0.0008	0.0008	0.0314

Table 6.20 P-values from additive models for total cancer mortality with exclusion of years of follow-up

The results remained similar when the early years of follow-up were excluded, with the exception of cholesterol which became statistically non-significant after excluding the first two years. It was possible that this was simply due to a widening of the confidence intervals for the difference between the levels of cholesterol, caused by the decrease in the number of subjects included in the analysis. For each of the three models fitted, the lowest cholesterol risk was associated with quintile 4, and the highest risk was associated with quintile 1 of cholesterol levels. The differences $(\hat{\beta}_1 - \hat{\beta}_4)$ and the standard errors $(\sqrt{\text{var}(\hat{\beta}_1 - \hat{\beta}_5) + \text{var}(\hat{\beta}_4 - \hat{\beta}_5) - 2\text{cov}(\hat{\beta}_1 - \hat{\beta}_5, \hat{\beta}_4 - \hat{\beta}_5)})$ for this comparison are given in Table 6.21.

Model	Effect	Standard error
Full follow-up	0.4801	0.0914
Year 1 excluded	0.4014	0.1000
Years 1 and 2 excluded	0.3181	0.1171

Table 6.21 Differences and standard errors for comparison of cholesterol quintiles 1 and 4 for each model

It has been stated elsewhere (Salmond et al, 1985) that if the magnitudes of relative risk estimates are not reduced when deaths in the first few years of follow-up are excluded then it is unlikely that cancer causes low cholesterol shortly before diagnosis. The odds ratios observed here for total cancer mortality for cholesterol quintile 1 to quintile 4 were:- 1.62 when all data was included

1.49 when the first year of follow-up was excluded

1.37 when the first two years of follow-up were excluded.

These odds ratios showed a trend consistent with a reduced effect, as the early years of follow-up were excluded, but this decrease cannot be confirmed due to the large standard errors (Table 6.21).

6.3.1.6 Exclusion of competing causes of death

A further point of interest was to see whether the relationships between the risk factors and the outcome of total cancer mortality were strengthened by excluding all deaths from non-cancer causes, as described in section 6.2.4. The resulting p-values are shown in Table 6.22.

Risk factor	All subjects	Excluding non-cancer deaths
Age	<0.0001	<0.0001
Smoking	<0.0001	<0.0001
DBP	0.9645	0.9136
BMI	<0.0001	<0.0001
Cholesterol	<0.0001	<0.0001
Alcohol	0.1165	0.0626
DEPCAT	0.0008	0.0003

Table 6.22 P-values for multivariate main effects models for total cancer mortality with exclusion of competing causes of death.

The results remained very similar after excluding these competing causes of death. The only factor on which this exclusion had any appreciable effect was alcohol consumption. When all 73,110 men were included in the analysis, alcohol consumption had a p-value of 0.1165 and there were no significant pairwise differences (with Bonferroni correction) between levels of alcohol. However, when men dying from non-cancer causes were excluded from the model fitting, there was a significant difference between levels 1 and 2 of alcohol (that is, 0 units and 1-10 units) with men consuming 0 units per week being at significantly higher risk of dying from cancer than men consuming 1-10 units of alcohol per week. The overall p-value for the alcohol effect was not quite significant though (p=0.0626), and this was really a very small effect.

6.3.2 Discussion

Middle-aged men were targeted for screening (as described in section 2.2) with around 80,000 attending for screening, although there was only an approximately 50% response rate with around 160,000 men being invited by letter. They were the only group screened in a systematic fashion. Consideration of this group of men thus gives a study of the people who are likely to come along to public health screening.

For mortality, the older age group was at increased risk for each of the Causes of Death considered (all-cause, CHD, cancer and trauma). Smoking also, predictably, was related to increased mortality risk from each of the major outcomes, although not from trauma. High blood pressure was associated with CHD mortality risk but not risk of cancer or trauma. Low BMI was related to increased risk of cancer and all-cause mortality, possibly due to weight loss induced by the cancer process. Cholesterol was a significant risk factor for each of the outcomes except trauma, for which it was only borderline significant, probably due to the small number of trauma events observed. Cholesterol showed the same trends reported elsewhere (see chapter 1) - the U-shaped curve for all-cause mortality being reflected by the lowest and highest quintiles being at significantly higher risk than the middle three quintiles. The risk in the highest quintile was associated with coronary causes, with risk of CHD increasing in order from quintile 1 up

to quintile 5. The risk in the lowest quintile was mainly associated with cancer. Alcohol was significant for each outcome except cancer. Trauma risk increased with alcohol consumption, but the highest risk of CHD mortality was for the group consuming 0 units of alcohol per week. DEPCAT also showed interesting relationships with the major causes of death - the most deprived group being at the highest risk. These results (see Table 6.7) are in agreement with the findings of other studies, as discussed in Chapter 1. However, the main effects model becomes difficult to interpret when significant interaction terms are included in the model.

The greatest risks of hospitalisation from any of the outcomes were for the older age group, the current smokers, and the deprived group (see Table 6.14). High BMI was related to increased risk of CHD, but lower risk of cancer. Cholesterol also remained consistent to the mortality result already seen - highest all-cause risk for quintiles 1 and 5, graded relationship with CHD with highest risk in quintile 5, and highest cancer risk in quintile 1. Alcohol showed the expected pattern with trauma hospitalisation, that is higher risk with higher consumption, but the CHD risk was significantly higher for consumption of 0 units per week. For hospitalisation from any cause, the categories of 0 and >60 units of alcohol per week were at significantly higher risk than the intermediate categories. These results are thus similar to those obtained for mortality in 45-64 year old men.

The final models arrived at in this study were identified using forward stepwise logistic regression techniques, which seem to have given plausible results. However, it is possible that a backward stepwise or an all-subsets method could have resulted in different models which fitted the data equally well. The main reason for sticking with the forward stepwise method was the reduction in computer run time, avoidance of problems with array sizes (for storage of the design matrix), and simplicity. The models which were identified using the forward stepwise method were checked for fit by means of a plot of residuals against fitted values and examination of the dispersion parameters. These checks revealed no problems in model fit. Only one example of the resulting residual plots and dispersion parameters is given in this thesis (see Appendix H).

For the comparison of cancer events on the three types of record, the main differences among the results observed were for cholesterol and alcohol. Cholesterol was highly significant for the deaths and hospitalisations, but not for the SMR6s. This may be due to the lower number of incidence records for the SMR6s which only go up to 1992, but there was still a larger number of SMR6s until 1992 than there were deaths until 1993, so this explanation seems unlikely. It is possible that cholesterol may be more strongly related to the more serious cancers requiring hospital treatment and frequently causing death, whilst having a weaker relationship when 'minor' skin neoplasms are included in the analysis. A further plausible explanation would be that cholesterol lowering is more apparent with 'advanced' cancers, as appearing in the deaths and hospitalisations, but less clear for cancers which have been diagnosed recently, as documented on cancer registration forms (SMR6s). This would provide some support for the hypothesis that cholesterol lowering is an effect of existing cancer and not a causal effect in itself. Alcohol was not a significant risk factor for cancer mortality, but was for cancer incidence using either SMR1s or SMR6s. As for mortality, the highest cancer incidence risk was for the categories of >50 or 0 units of alcohol per week. This difference in results could be explained by the cumulation of more events for incidence than for mortality, enabling better detection of significant effects.

The exclusion of early years of follow-up had little effect on the relationships between the risk factors and cancer mortality, with the exception of cholesterol. Cholesterol became marginally non-significant after the exclusion of the first two years of follow-up. While the difference between the lowest (quintile 4) and the highest (quintile 1) risk groups decreased as years were excluded, the standard error for this difference increased slightly. The odds ratios for mortality for quintile 1 to quintile 4 showed a downward trend as years were excluded, but the excess risk in quintile 1 relative to quintile 4 had not disappeared after excluding the first two years. This supports the findings of other studies (see section 1.4) which have suggested that the relationship between cancer and low cholesterol may be a short term effect of undiagnosed disease which disappears once the first few years of follow-up are excluded. Although insufficient years have been excluded to remove the apparent effect in this study, the trend is heading that way and this should be examined again once more years of follow-up have been accumulated.

It is interesting to note that the exclusion of competing causes of death had little effect on the results obtained from the logistic regression models. Alcohol came close to overall significance for cancer mortality when alternative causes of death were excluded, so this may be worth re-examining once further years of follow-up have been accumulated.

6.4 Women who attended screening visit 1

6.4.1 Results

The mean period of follow-up for women who attended screening visit 1 was 4.6 years (68646 subject years in total) until the end of 1993. The deaths identified by record linkage for these 14,950 women are given in Table 6.23, broken down by cause of death.

Cause of death	Number of deaths
Malignant neoplasms	155
Coronary heart disease	86
Cerebrovascular	33
Other circulatory	29
Respiratory	13
Digestive	8
Injury / poisoning	6
Other cause	8
Total	**338**

Table 6.23 Numbers of deaths for women screened

Table 6.24 contains the frequencies and rates of death for the 14,950 women screened broken down by age group (<45, 45-54, 55-64 or >64 years), smoking status (never smoked, ex-smoker or current smoker), diastolic blood pressure (≤ or > the median pressure of 80 mmHg), body mass index (< or ≥ the median of 24.5 kg/m^2), cholesterol (quintiles - <4.97, 4.97-5.67, 5.68-6.33, 6.34-7.16 or >7.16 mmol/l), alcohol (0, 1-10 or >10 units per week) and DEPCAT (affluent, middle or deprived). The death rates are calculated based on the number of subject years of follow-up accrued in each level of each factor.

Variable	Levels	Number of subjects	Deaths (All-causes)	Deaths (CHD)	Deaths (Cancer)	Deaths (Injury / poisoning)
Age						
(years)	<45	3422	10 (0.64)	0 (0.00)	7 (0.45)	0 (0.00)
	45-54	4950	48 (2.12)	7 (0.31)	24 (1.06)	1 (0.04)
	55-64	4818	144 (6.50)	43 (1.94	65 (2.93)	4 (0.18)
	>64	1749	136 (16.83)	36 (4.45)	59 (7.30)	1 (0.12)
Smoking						
	Never	7976	135 (3.68)	33 (0.90)	61 (1.66)	4 (0.11)
	Ex-smoker	2927	80 (5.94)	21 (1.56)	38 (2.82)	0 (0.00)
	Current	4045	123 (6.64)	32 (1.73)	56 (3.02)	2 (0.11)
Cholesterol						
(mmol/l)	<4.97	2885	32 (2.44)	4 (0.30)	17 (1.30)	0 (0.00)
	4.97-5.67	2960	68 (5.04)	12 (0.89)	32 (2.37)	1 (0.07)
	5.68-6.33	3016	71 (10.24)	18 (2.60)	32 (4.62)	1 (0.14)
	6.34-7.16	3012	66 (4.76)	21 (1.51)	30 (2.16)	0 (0.00)
	>7.16	3069	101 (7.07)	31 (2.17)	44 (3.08)	4 (0.28)
Alcohol						
(units/week)	0	6750	193 (6.19)	53 (1.70)	83 (2.66)	3 (0.10)
	1-10	7118	126 (3.87)	30 (0.92)	63 (1.94)	2 (0.06)
	>10	753	12 (3.51)	2 (0.58)	6 (1.75)	0 (0.00)
DBP						
(mmHg)	≤ 80	8653	169 (4.26)	41 (1.03)	77 (1.94)	4 (0.10)
	> 80	6291	169 (5.84)	45 (1.55)	78 (2.69)	2 (0.07)
BMI						
(kg/m^2)	< 24.5	7539	170 (4.91)	37 (1.07)	83 (2.40)	2 (0.06)
	≥ 24.5	7398	167 (4.92)	48 (1.41)	72 (2.12)	4 (0.12)
DEPCAT						
	Affluent	1536	23 (3.28)	6 (0.86)	9 (1.28)	1 (0.14)
	Middle	10529	255 (5.24)	65 (1.33)	124 (2.54)	2 (0.04)
	Deprived	2805	60 (4.76)	15 (1.19)	22 (1.75)	3 (0.24)

Table 6.24: Number of deaths (events/1000 years of follow-up) for women attending screening visit 1 broken down by each of the risk factors. It should be noted that due to a small amount of missing data for each variable the total numbers of subjects are not consistent across the variables.

The women show lower mortality rates than the middle-aged men considered in section 6.3. In many of the categories, the female mortality rate is approximately half that in the corresponding category for middle-aged men. On the whole, the trends appear similar to those observed for men.

Main effects models

Logistic regression main effects models were fitted for these categorised deaths. There were 14,517 women who had no missing values on the variables contributing to the category and could thus be used in fitting models. The p-values obtained when each variable was fitted on its own are given in Table 6.25, and the p-values obtained for each variable once all the other factors in the category were included are given in Table 6.26.

Risk factor	All-cause mortality	Cancer mortality	CHD mortality
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	0.0048	0.0275
DBP	0.0033	0.0317	0.0564
BMI	0.0219	0.4905	0.2364
Cholesterol	<0.0001	0.0481	0.0015
Alcohol	<0.0001	0.1076	0.0083
DEPCAT	0.0310	0.0221	0.4139

Table 6.25 P-values for univariate logistic models for female mortality

Risk factor	All-cause mortality	Cancer mortality	CHD mortality
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	0.0007	0.0018
DBP	0.3383	0.2985	0.6268
BMI	0.2420	0.1768	0.6436
Cholesterol	0.0675	0.4381	0.6684
Alcohol	0.0711	0.7313	0.2378
DEPCAT	0.0837	0.0344	0.5754

Table 6.26 P-values for multivariate main effects logistic models for female mortality. P-values are for each variable once all the other factors have already been fitted.

The pairwise contrasts between the levels of each factor (Bonferroni corrected) gave an indication of the direction of increased mortality risk with results as in Table 6.27.

Risk factor	All-cause mortality	Cancer mortality	CHD mortality
Age	1 2 3 4	1 2 3 4	1 2 3 4
Smoking	1 2 3	1 2 3	1 2 3
DBP	1 2	1 2	1 2
BMI	2 1	2 1	1 2
Cholesterol	4 1 5 3 2	4 5 3 1 2	1 4 2 3 5
Alcohol	2 3 1	3 2 1	3 2 1
DEPCAT	1 3 2	1 3 2	3 1 2

Table 6.27 Significance of pairwise contrasts in multivariate logistic main effects models for female mortality. Levels of each risk factor are given in order of increasing risk. Levels which are joined by a line were not significantly different from each other after adjusting for multiple comparisons by means of a Bonferroni correction.

Stepwise modelling

The next stage of the analysis was to allow interaction terms to be brought into the logistic model, in a forward stepwise manner as has been described earlier.

For all-cause mortality, only the main effects of age and smoking were brought into the model. The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for all-cause mortality are given in Table 6.28. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F10 to examine any contrast between levels of the factors.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(45-54 / <45)	1.20	0.35	3.44	3.31
(55-64 / <45)	2.34	0.33	7.12	10.4
(>64 / <45)	3.4	0.33	10.4	31.2
smoking(ex/never)	0.32	0.15	2.19	1.37
(current/never)	0.80	0.13	6.08	2.22
CONSTANT	-6.16	0.32	-18.9	0.00

Table 6.28 Parameter estimates for the contrasts in the final model for all-cause mortality in the women screened. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

For cancer mortality, the main effects of age, smoking and DEPCAT were brought into the model. The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for cancer mortality are given in Table 6.29. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F11 to examine any contrast between levels of the factors.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(45-54 / <45)	0.87	0.43	2.02	2.38
(55-64 / <45)	1.92	0.40	4.80	6.79
(>64 / <45)	2.85	0.40	7.06	17.3
smoking(ex/never)	0.38	0.21	1.83	1.47
(current/never)	0.74	0.19	3.89	2.11
depcat(middle/affluent)	0.72	0.37	1.97	2.06
(deprived/affluent)	0.29	0.42	0.68	1.33
CONSTANT	-7.10	0.52	-13.6	0.00

Table 6.29 Parameter estimates for the contrasts in the final model for cancer mortality in the women screened. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

For CHD mortality, age, smoking and the age*smoking interaction were brought into the model. The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for CHD mortality are given in Table 6.30. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F12 to examine any contrast between levels of the factors.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(45-54 / <45)	2.38	2.54	0.94	10.8
(55-64 / <45)	4.50	2.37	1.90	89.7
(>64 / <45)	6.52	2.35	2.77	680.
smoking(ex/never)	0.00	4.98	0.00	1.00
(current/never)	0.00	3.92	0.00	1.00
Age45-54(Ex/Never smoker) / Age<45(Ex/Never smoker)	1.065	5.18	0.206	2.90
Age55-64(Ex/Never smoker) / Age<45(Ex/Never smoker)	1.26	5.00	0.25	3.51
Age>64(Ex/Never smoker) / Age<45(Ex/Never smoker)	-0.39	5.00	-0.08	0.67
Age45-54(Current/Never smoker) / Age<45(Current/Never smoker)	2.15	4.07	0.53	8.59
Age55-64(Current/Never smoker) / Age<45(Current/Never smoker)	1.66	3.94	0.42	5.27
Age>64(Current/Never smoker) / Age<45(Current/Never smoker)	-0.31	3.95	-0.08	0.73
CONSTANT	-10.20	2.34	-4.35	0.00

Table 6.30 Parameter estimates for the contrasts in the final model for CHD mortality in the women screened. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

The observed proportions of deaths at each level of the interaction are tabulated in Table 6.31, and the expected proportions are tabulated in Table 6.32.

Observed	Never smoker	Ex-smoker	Current smoker
Age <45	0.0000	0.0000	0.0000
Age 45-54	0.0004	0.0012	0.0034
Age 55-64	0.0032	0.0115	0.0170
Age >64	0.0236	0.0166	0.0176

Table 6.31 Observed risk of death for each level of age crossed with smoking

Expected	Never smoker	Ex-smoker	Current smoker
Age <45	0.0000	0.0000	0.0000
Age 45-54	0.0010	0.0014	0.0023
Age 55-64	0.0061	0.0086	0.0143
Age >64	0.0160	0.0226	0.0371

Table 6.32 Expected risk of death for each level of age crossed with smoking, based on the logistic main effects model for CHD mortality

The proportion of deaths / subjects increased with age for each smoking category, and increased with smoking category for each age group except age <45 in which there were no deaths. The nature of the interaction will be considered further in the discussion (section 6.4.2).

6.4.2 Discussion

The women screened represent a more diverse population than the men considered earlier. They were not invited to screening, but they were included since screening was initially open to everyone. However, in the later stages of the screening process, women were not included. Many of the women screened had accompanied their husbands to their first screening visit. However, despite the difficulties in clearly defining this population, the women screened are of great interest since most of the previous studies (see Chapter 1) have concentrated on middle-aged men, with relatively few of them considering women. For these women, there was only a small number of deaths, with the proportion of subjects dying being only approximately half that observed in the middle-aged men.

While many variables appear to be significantly related to each of the outcomes at a univariate level (see Table 6.25), once all the risk factors are included only age and smoking remain significant (see Table 6.26), although other factors showed patterns which were similar to those for men (see Table 6.27). For example, the highest risk (for each outcome) is for an alcohol consumption of 0 units per week, although alcohol was

not significant overall for any outcome. However, it should be noted that 46% of the women had a reported alcohol consumption of 0 units per week. For the women, there was a wide age range, and age-specific mortality rates for women tend to be lower than those for men, so that insufficient events were observed for this cohort of women to clearly detect relationships between baseline risk factors and mortality. The pairwise contrasts may also be affected by the differing numbers of women in the various levels. For example, the majority of women were in level 2 of DEPCAT and this may lead to difficulty in detecting differences between levels 1 and 3. However, it is clear that the most outstanding modifiable risk factor for women was smoking, with cholesterol level being of much less importance. Cholesterol was almost significant for all-cause mortality, with quintile 2 being at significantly higher risk than quintile 4, but it was far from significant for any of the more specific outcomes. It is possible that, for women, cholesterol may be partially confounded with age. The only significant interaction term was age*smoking, for CHD mortality, again in agreement with the recent case-control study (Parish et al, 1995) which was discussed in section 6.3.2. These two interacting variables had multiplicative effects. None of the contrasts estimated to help interpret this interaction came close to significance (see Table 6.30), so this significant overall interaction may be due to sensitivity in picking up very small effects. Each of the models was checked by means of a plot of residuals against fitted values and consideration of the dispersion parameters (as described in Appendix H). There were no indications that there was a problem with model fit.

Consideration of incidence may yield higher event rates, and more power to detect differences. However the low event rate for CHD is due to the relatively young age of the women, since, in general, CHD occurs at an older age in women than men (see section 1.1.2), so that consideration of incidence is unlikely to yield a sufficient frequency of CHD events. More events will be accumulated with further years of follow-up and an ageing cohort.

6.5 Men who reached screening visit 3

6.5.1 Results

The deaths identified by record linkage for the 13,559 men who reached screening visit 3 are given in Table 6.33.

Cause of death	Number of deaths
Malignant neoplasms	135
Coronary heart disease	205
Cerebrovascular	23
Other circulatory	11
Respiratory	17
Digestive	14
Injury / poisoning	18
Other cause	39
Total	**462**

Table 6.33 Numbers of deaths for men who reached screening visit 3

Table 6.34 contains the frequencies and rates of death for the 13,559 men who reached screening visit 3 broken down by age (45-54 or 55-64 years), smoking status (never smoked, ex-smoker or current smoker), diastolic blood pressure (< or ≥ the median pressure of 85 mmHg), LDL cholesterol (< or ≥ the median of 4.75 mmol/l), HDL cholesterol (tertiles - <1.50, 1.05-1.2 or >1.2 mmol/l), fibrinogen (quintiles - <3.59, 3.59-3.99, 4.0-4.40, 4.41-4.92, or >4.92 g/dl) and viscosity (tertiles - <1.29, 1.29-1.37 or >1.37 mPa.s). The death rates are calculated based on the number of subject years of follow-up accrued within each level of each factor.

Variable	Levels	Number of subjects	Deaths (All-causes)	Deaths (CHD)	Deaths (Cancer)	Deaths (Injury / poisoning)
Age (years)	45-54	6784	122 (4.47)	61 (2.24)	24 (0.88)	10 (0.37)
	55-64	6653	336 (12.34)	142 (5.21)	109 (4.00)	8 (0.29)
Smoking						
	Never	3720	67 (4.47)	29 (1.93)	18 (1.20)	2 (0.13)
	Ex-smoker	3999	113 (6.94)	52 (3.19)	30 (1.84)	3 (0.18)
	Current	4719	282 (11.94)	124 (5.25)	87 (3.68)	13 (0.55)
DBP (mmHg)	≤ 85	6785	193 (7.41)	71 (2.73)	69 (2.65)	9 (0.34)
	> 85	6743	269 (9.32)	134 (4.64)	66 (2.29)	9 (0.31)
LDL chol (mmol/l)	< 4.75	6408	206 (7.89)	90 (3.45)	67 (2.57)	8 (0.31)
	≥ 4.75	7120	256 (8.89)	115 (3.99)	68 (2.36)	10 (0.35)
HDL chol (mmol/l)	< 1.05	4769	190 (9.88)	96 (4.99)	58 (3.01)	6 (0.31)
	1.05-1.2	3180	150 (7.52)	65 (3.26)	37 (1.85)	7 (0.35)
	>1.2	5579	122 (7.77)	44 (2.80)	40 (2.55)	5 (0.32)
Viscosity (mPa.s)	< 1.29	5298	106 (4.97)	51 (2.39)	30 (1.41)	4 (0.19)
	1.29-1.37	4858	147 (8.24)	57 (3.20)	48 (2.69)	8 (0.45)
	>1.37	3372	191 (12.13)	88 (5.59)	54 (3.43)	5 (0.32)
Fibrinogen (g/dl)	<3.59	2963	62 (5.01)	26 (2.10)	24 (1.94)	2 (0.16)
	3.59-3.99	2581	57 (5.39)	24 (2.27)	14 (1.32)	4 (0.38)
	4.0-4.40	2686	78 (6.94)	35 (3.11)	24 (2.13)	2 (0.18)
	4.41-4.92	2630	91 (8.82)	45 (4.36)	25 (2.42)	4 (0.39)
	>4.92	2668	161 (15.46)	68 (6.53)	46 (4.42)	6 (0.58)

Table 6.34: Number of deaths (events/1000 years of follow-up) for men who reached screening visit 3 broken down by each of the risk factors. It should be noted that due to a small amount of missing data for each variable the total numbers of subjects are not consistent across the variables.

The men who reached screening visit 3 had lower mortality rates than the large cohort of men at screening visit 1 considered earlier. The rates for this sub-cohort were approximately two thirds of those observed for all middle-aged men for all-cause mortality, and about half those observed for cancer mortality. The rates were similar, although still slightly lower, for mortality from CHD or trauma.

Main effects models

Logistic main effects models were fitted for these categorised deaths. There were 12,864 men who had no missing values on the variables contributing to the category and could thus be used in fitting models. The p-values obtained when each variable was fitted on its own are given in Table 6.35, and the p-values obtained for each variable once all the other factors were included are given in Table 6.36.

Risk factor	All-cause mortality	Cancer mortality	CHD mortality
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	<0.0001	<0.0001
Fibrinogen	<0.0001	0.0006	<0.0001
HDL	0.0089	0.0528	0.0018
LDL	0.2509	0.6201	0.3083
Viscosity	<0.0001	0.0016	<0.0001
DBP	0.0050	0.3985	<0.0001

Table 6.35 P-values for univariate logistic models for mortality in men who reached screening visit 3

Risk factor	All-cause mortality	Cancer mortality	CHD mortality
Age	<0.0001	<0.0001	<0.0001
Smoking	<0.0001	<0.0001	<0.0001
Fibrinogen	0.0094	0.1595	0.3888
HDL	0.1943	0.1502	0.0305
LDL	0.9603	0.2591	0.7191
Viscosity	0.0033	0.0573	0.0855
DBP	0.0094	0.3342	0.0001

Table 6.36 P-values for multivariate main effects logistic models for mortality in men who reached screening visit 3. P-values are for each variable once all the other factors have already been fitted.

The pairwise contrasts between the levels of each factor (Bonferroni corrected) gave an indication of the direction of increased mortality risk, with results as in Table 6.37.

Risk factor	All-cause mortality	Cancer mortality	CHD mortality
Age	1 2	1 2	1 2
Smoking	1 2 3	1 2 3	1 2 3
Fibrinogen	2 3 4 1 5	2 4 3 5 1	2 1 3 4 5
HDL	2 3 1	2 3 1	3 2 1
LDL	2 1	2 1	1 2
Viscosity	1 2 3	1 2 3	1 2 3
DBP	1 2	2 1	1 2

Table 6.37 Significance of pairwise contrasts in multivariate main effects logistic models for mortality. Levels of each risk factor are listed in order of increasing risk. Levels which are joined by a line were not significantly different from each other after adjusting for multiple comparisons by means of a Bonferroni correction.

Stepwise modelling

The next stage of the analysis was to allow interaction terms to be brought into the logistic model, in a forward stepwise manner as has been described earlier.

For all-cause mortality, the final model resulting from this forward stepwise process dropped the HDL and LDL main effects and introduced the age*smoking interaction (p=0.038).

The parameter estimates, approximate z-statistics (coefficient / standard error) and odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for all-cause mortality are given in Table 6.38. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F13 to examine any contrast between levels of the factors.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	1.03	0.11	9.27	2.81
smoking(ex/never)	0.64	0.40	1.63	1.91
(current/never)	1.44	0.37	3.90	4.21
viscosity(medium/low)	0.31	0.14	2.27	1.36
(high/low)	0.49	0.14	3.53	1.64
fibrinogen(q2/q1)	-0.35	0.50	-0.71	0.70
(q3/q1)	0.19	0.43	0.45	1.21
(q4/q1)	0.62	0.41	1.51	1.86
(q5/q1)	1.04	0.41	2.56	2.83
dbp(high/low)	0.26	0.10	2.56	1.29
Ex-smoker(q2/q1) / Never smoker(q2/q1)	0.11	0.61	0.18	1.12
Ex-smoker(q3/q1) / Never smoker(q3/q1)	0.17	0.56	0.30	1.18
Ex-smoker(q4/q1) / Never smoker(q4/q1)	-0.65	0.56	-1.16	0.52
Ex-smoker(q5/q1) / Never smoker(q5/q1)	-0.22	0.49	-0.44	0.81
Current smoker(q2/q1) / Never smoker(q2/q1)	-0.49	0.52	-0.95	0.61
Current smoker(q3/q1) / Never smoker(q3/q1)	-0.91	0.47	-1.93	0.40
Current smoker(q4/q1) / Never smoker(q4/q1)	-0.46	0.49	-0.94	0.63
Current smoker(q5/q1) / Never smoker(q5/q1)	-0.97	0.45	-2.14	0.38
CONSTANT	-5.37	0.33	-16.0	0.00

Table 6.38 Parameter estimates for the contrasts in the final model for all-cause mortality in males age 45-64 at screening visit 3. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

The observed proportions (deaths/subjects) are tabulated in Table 6.39. The risk of death increased with smoking category for each level of fibrinogen, but showed no clear trend with fibrinogen in each smoking category. The expected proportions are given in Table 6.40.

Fibrinogen level	Never smoked	Ex-smoker	Current smoker
0-3.58 g/dl	0.0105	0.0215	0.0421
3.59-3.99 g/dl	0.0084	0.0190	0.0375
4.0-4.40 g/dl	0.0162	0.0170	0.0463
4.41-4.92 g/dl	0.0271	0.0329	0.0395
>4.92 g/dl	0.0447	0.0589	0.0663

Table 6.39 Observed risk of death for each level of fibrinogen crossed with smoking

Fibrinogen level	Never smoked	Ex-smoker	Current smoker
0-3.58 g/dl	0.0179	0.0249	0.0462
3.59-3.99 g/dl	0.0151	0.0210	0.0377
4.0-4.40 g/dl	0.0175	0.0243	0.0436
4.41-4.92 g/dl	0.0183	0.0254	0.0455
>4.92 g/dl	0.0266	0.0364	0.0635

Table 6.40 Expected risk of death for each level of fibrinogen crossed with smoking, based on the logistic main effects model for all-cause mortality

For cancer mortality, the main effects of age, smoking and viscosity were brought into the model. The parameter estimates, approximate z-statistics (coefficient / standard error) and estimated odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for all-cause mortality are given in Table 6.41. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F14 to examine any contrast between levels of the factors.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64/45-54)	1.58	0.23	6.82	4.86
smoking(ex/never)	0.36	0.30	1.19	1.44
(current/never)	1.13	0.27	4.19	3.08
viscosity(medium/low)	0.49	0.24	2.07	1.64
(high/low)	0.60	0.23	2.54	1.81
CONSTANT	-6.7	0.34	-19.7	0.00

Table 6.41 Parameter estimates for the contrasts in the final model for cancer mortality in males age 45-64 at screening visit 3. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

For CHD mortality, age, smoking, DBP, viscosity and HDL were brought into the model. The parameter estimates, approximate z-statistics (coefficient / standard error) and estimated odds ratios (exponential(coefficient)) obtained by fitting the final logistic model for all-cause mortality are given in Table 6.42. These parameter estimates may be used in conjunction with the covariance matrix given in Appendix F15 to examine any contrast between levels of the factors.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64/45-54)	0.84	0.16	5.29	2.31
smoking(ex/never)	0.47	0.24	1.94	1.60
(current/never)	1.03	0.22	4.70	2.81
dbp(high/low)	0.63	0.16	4.06	1.88
viscosity(medium/low)	0.10	0.20	0.52	1.11
(high/low)	0.54	0.18	2.97	1.71
HDL chol(medium/low)	-0.35	0.17	-2.07	0.71
(high/low)	-0.49	0.19	-2.58	0.61
CONSTANT	-5.71	0.28	-20.1	0.00

Table 6.42 Parameter estimates for the contrasts in the final model for CHD mortality in males age 45-64 at screening visit 3. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

Each of the models was checked by means of a plot of residuals against fitted values, and examination of the dispersion parameters (as described in Appendix H). There were no indications of a problem with model fit.

6.5.2 Discussion

The men who reached screening visit 3 represented a more homogeneous population than those who attended screening visit 1. The men who were invited to screening visit 1 were identified on the basis of age and sex. By the time they got to screening visit 3, this subgroup of men had additional characteristics. These included no history of symptomatic myocardial infarction, a raised total cholesterol level ($\geq$ 6.5 mmol/l at screening visit 1), and an elevated LDL cholesterol level (($\geq$ 4.0 mmol/l at screening visit 2). Further details of the screening process can be found in Chapter 2.

The main interest in this analysis was in the laboratory measurements of fibrinogen and plasma viscosity which other recent studies have shown to be potentially important risk factors for CHD (see section 1.1.4). Both of these variables were significantly related to each of the three outcomes (all-cause, CHD and cancer mortality) when considered at a univariate level. However, once all other main effect terms were included they were only significant for all-cause mortality. The univariate pairwise contrasts revealed that risk of any of the outcomes increased with plasma viscosity. Fibrinogen showed patterns similar to cholesterol in that the highest cancer risk was in quintile 1 and the highest CHD risk was in quintile 5. It was worthy of note that LDL cholesterol was not significantly related to any of the outcomes, and HDL cholesterol was related only to CHD mortality, for which the highest risk was with the lowest HDL levels. The only interaction term which was introduced by forward stepwise modelling was smoking*fibrinogen for all-cause mortality. The only significant contrast estimated to help interpret this interaction was for the difference between quintiles 1 and 5 of fibrinogen for current and never smokers (see Table 6.38). The contrast between quintiles 1 and 3 of fibrinogen for current and never smokers was close to significance. The proportion of deaths in the highest fibrinogen quintile was over four times that in

the lowest quintile for never smokers, but was only 1.6 times that in the lowest quintile for current smokers (see Table 6.39). Fibrinogen level thus seems to be a much more important factor in predicting all-cause mortality in the absence of smoking. However, this interaction could be an artefact of the varying distribution of fibrinogen levels in smokers and non-smokers. While fibrinogen levels are split into quintiles, the number of subjects in each cell when these quintiles are cross-tabulated by smoking category show differing patterns (see Table 6.43). The mean and median values in the highest quintile (which has no upper limit) are also slightly higher with higher levels of smoking (see Table 6.44), which may provide another explanation of the significant interaction effect.

Fibrinogen quintile	Never smoker	Ex-smoker	Current smoker
<3.59	998	913	764
3.59-3.99	861	820	914
4.0-4.40	769	858	1166
4.41-4.92	607	684	1304
>4.92	424	658	1569

Table 6.43. Numbers of subjects at levels of fibrinogen*smoking.

Fibrinogen quintile	Never smoker	Ex-smoker	Current smoker
<3.59	3.23 / 3.32	3.23 / 3.31	3.23 / 3.34
3.59-3.99	3.80 / 3.81	3.80 / 3.81	3.80 / 3.81
4.0-4.40	4.19 / 4.18	4.20 / 4.20	4.20 / 4.20
4.41-4.92	4.64 / 4.63	4.66 / 4.67	4.67 / 4.66
>4.92	5.57 / 5.36	5.62 / 5.40	5.63 / 5.46

Table 6.44. Mean/Median values of fibrinogen in each cell of fibrinogen*smoking.

6.6 Overview

This prospective follow-up study raised a number of ethical issues. If individual patient records are to be accessed by study investigators, it is important that full informed consent is obtained from the trial participants prior to the study. Ideally, the patients should be aware that, by giving an investigator access to their medical records, the investigator will be able to access both hardcopy and computerised versions of their records. In the epidemiological follow-up of WOSCOPS screenees there was no full informed consent from the screenees to access their medical records. The data which have been accessed by record linkage are routinely published in summary form and are likewise made available to bona fide investigators in the form of small-area statistics, all without the permission of the individual patients. It is argued that use of the Scottish Record Linkage System to link screening data to national databases of medical records, in order to obtain output in the form of summary statistics, does not breach any rules of confidentiality and is comparable with routine uses for such data. Investigators should be aware that access will be denied to certain sensitive medical information even if informed consent is available. For example, in this study no access was given to any information which would reveal the HIV status of the subjects. For linkages carried out at the Scottish Record Linkage System, each study is considered in detail by a Privacy Advisory Committee and must be approved by them before any linkages can be carried out.

The analysis involving CHD risk factors in men age 45-64 confirmed that WOSCOPS was in agreement with the trends reported in other studies (see Chapter 1). It was reassuring to note that smoking and increasing age were found to be related to increased risk of all-cause, CHD and cancer mortality or hospitalisation. The most socially deprived group, as defined by DEPCAT groupings, were also at increased risk of these outcomes. The analysis reported in section 6.3 was repeated using systolic instead of diastolic blood pressure, yielding similar results. Cholesterol showed the U-shaped relationship with all-cause mortality, caused by high CHD risk with high cholesterol and high cancer risk with low cholesterol which has been reported elsewhere (Sorlie and Feinleib, 1982; Blackwelder et al, 1980). It was also interesting to identify a U-shaped

risk curve with alcohol consumption, with the highest risk category being 0 units per week, which has also been reported by other studies (Friedman and Kimball, 1986; Marmot et al, 1981). Analysis of the WOSCOPS cohort found that the association between low cholesterol and cancer mortality diminished with time and became non-significant when the first two years of follow-up were excluded. The odds ratio decreased from 1.62 to 1.37 when the first two years were excluded. It would thus seem likely that this relationship is due, at least in part, to the cholesterol lowering effects of a developing malignancy.

The results for women provided some interesting contrasts with those for men. For all-cause mortality, only age and smoking were significant for women, while all seven risk factors were significant for the middle-aged men. Similarly, for CHD, only age and smoking were significant for female mortality, while all except BMI were significant for the men. For cancer mortality, age, smoking and DEPCAT were significant for women, while age, smoking, DEPCAT, BMI and cholesterol were significant for men. Some of these differences in results may be due to the fact that, for men, analysis was restricted to the 45-64 year age group, while, for women, the full range of ages were included. The differences may also be due to the fact that, in general, women have lower age-specific death rates than men, and thus insufficient numbers of events to detect differences have been observed for women, and, probably more importantly, many fewer women were studied.

The results for the men who reached screening visit 3 provide an interesting supplement to the earlier results, with the WOSCOPS results reflecting those in other studies.

Obviously there is huge potential for analyses of alternative diagnoses for the screened cohort. While there has been insufficient length of follow-up to detect some effects, this preliminary analysis gives clear indications of factors of interest for future work.

The main restriction on the analysis so far has been the use of categorised data. While many variables can be examined in this way, potential cross-classifications have been abandoned due to small cell sizes caused by correlations between variables. The full potential of this data set has thus not been utilised. Use of the original data set (with

continuous explanatory variables) in the analysis would enhance the power to detect associations between risk factors and outcome and allow consideration of combinations of variables which have been impossible so far. However, analysis using the original data set would involve identifying each patient and their event records with the same ID number, and thus break the requirements of subject anonymity.

An alternative to analysis by categorical versions of the baseline variables would be to construct a logistic score function from the logistic regression coefficients of models involving continuous explanatory variables. Since these coefficients are unavailable for the WOSCOPS screened cohort (NB individual subject data not available due to privacy restrictions), the alternative would be to use coefficients from models fitted in other published studies, for example, MRFIT or Framingham. A categorisation of this score could be used to adjust for some variables, and thus allow further factors to be examined in a categorical analysis, without compromising subject anonymity.
Problems with this method of analysis arise in:
1) choice of which published study to obtain logistic regression coefficients from. The study used should, ideally, relate to a population similar to the one currently under investigation.
2) which variables to adjust for in the score. It would be necessary to adjust for all the variables contributing to the published model, some of which may not be available for WOSCOPS, or some of which may be of interest in their own right. The published coefficients are affected by each variable in the model, and thus all should be included.
3) the outcome for the model fitted. The logistic score will be correlated with the particular outcome for the model from which the fitted coefficients came, and may not be appropriate for consideration of other outcomes.
This method may be worthy of further exploration in the future.

An alternative to logistic regression analysis would have been to fit some sort of survival curve for mortality against subject years of follow-up. This was impossible due to the format of the data - by category, and not identifiable at an individual subject level.

Chapter 7

Comparison of screenee and population event rates

7.1 Introduction

The relationship between event rates in the group of subjects who are willing to attend for health screening and that in the local community is of particular interest for the planning of public health initiatives. The main aim of this chapter is to compare event rates in the WOSCOPS screened cohort with those in the general population, in order to provide an assessment of how representative the screened cohort is of the population from which it comes. This will give an indication of the appropriateness of applying the results achieved for the screened cohort in chapter 6 to the general population. Consideration will once again focus on the 45-64 year old men since they make up the largest subgroup in the cohort screened, and were the only group screened in a systematic fashion.

7.2 Preliminary comparison of screenee and trial events to the Scottish population

Table 7.1 below contains the frequencies and rates of death for the WOSCOPS screened subjects, with follow-up until the end of 1993, for various ranges of ICD codes. For comparison, the corresponding death rates for the WOSCOPS randomised subjects and

for the whole of Scotland in 1991 (as given in the Registrar General's Annual Report for 1991) are also quoted. The year of 1991 was selected since it lay in the middle of the follow-up period and was the year in which the census was carried out. These rates relate to men age 45-64.

Type of event	ICD9 codes	Scottish death rates in 1991	Number of deaths for the 74,576 WOSCOPS screenees (death rates)	Number of deaths for the 6595 WOSCOPS randomised subjects (death rates)
All causes		11.72	3493 (11.70)	166 (6.85)
Cardiovascular	390-459	5.42	1668 (5.59)	86 (3.55)
- CHD	410-414	4.23	1333 (4.46)	74 (3.05)
- stroke	430-438	0.7	190 (0.64)	8 (0.33)
Malignant neoplasm	140-208	3.8	1230 (4.12)	57 (2.35)
- lung	162	1.39	474 (1.59)	20 (0.82)
External causes	E800-E999	0.6	131 (0.44)	6 (0.25)
- suicide	E950-E959	0.2	51 (0.17)	1 (0.04)
Respiratory	460-519	0.76	186 (0.62)	5 (0.21)
Digestive	520-579	0.46	144 (0.48)	4 (0.16)
Other causes		0.68	134 (0.45)	7 (0.29)

Table 7.1 Numbers of deaths and mortality rates for men age 45-64 broken down by various causes of death. The rates quoted are numbers of events per 1000 subject years.

There is broad agreement between the rates for the Scottish population and the WOSCOPS screenees across the range of causes of death considered, with rates for WOSCOPS randomised subjects being lower. It is clear that the death rates for the WOSCOPS randomised subjects, despite being selected as a high risk group for coronary heart disease, are lower than in the screened cohort as a whole, in each of the categories of cause of death, although these rates may not be statistically significantly

lower in all categories due to small numbers of events for some of the cause of death outcomes.

The preliminary data reported on the epidemiological prospective study of WOSCOPS screenees matches well with Scottish national death rates. It would be expected that subjects who are willing to attend for population screening will be more healthy than those in the general population. Subjects who are seriously ill and subjects in areas of social deprivation, where mortality rates are typically highest, are the least likely to attend health screening clinics. Acting in the opposite direction is the fact that screening was conducted in areas with high death rates relative to Scotland as a whole.

The low cause-specific death rates in the randomised subjects, as compared to the screenee group, serves as a reminder of the dangers of using population data to project event rates in clinical trials. Although inclusion criteria for the trial (including raised LDL cholesterol) would serve to select a group at higher risk for coronary heart disease, this would probably be compensated for by exclusion criteria (such as, no previous myocardial infarction, no recent evidence of coronary heart disease and no current serious illness). The danger of using population data to project event rates is particularly clear in the non-coronary death rates, where there would be little expectation of the trial medication lowering rates of death.

7.3 Population comparisons

Some initial comparisons were carried out to investigate the similarity between the population in the screening area and the Scottish population as a whole, and to look at variation within the screening area population.

7.3.1 Comparison of the screening area population with the Scottish national population

Mortality rates for the screening area population were calculated using data supplied by the Registrar General for Scotland. The screening area population was defined to be the general population resident in the postcode sectors in which screening was carried out. These postcode sectors are listed in Appendix I, and cover the four health board areas involved in the study - Greater Glasgow Health Board, Argyll and Clyde Health Board, Lanarkshire Health Board and Dumfries and Galloway Health Board. The Registrar General provided date of death and causes of death by sex, age and postcode sector, for all deaths occurring in the area of screening (as defined by postcode sectors) between 1988 and 1993. The Registrar General also provided the population size broken down by sex and 5-year age band based on the 1991 census. The relative frequency of deaths per annum for the screening area was calculated as x/6n where x was the number of deaths occurring in the 6-year period from 1988 until 1993, and n was the number of subjects according to the 1991 census.

The relative frequencies of deaths in the screening population and the whole of Scotland were compared for all-cause mortality in men age 45-64. The Scottish national death rate (as given in the Registrar General's Annual Report for 1991) was 11.7 deaths/1000 subject years. The all-cause mortality rate for the population of men age 45-64 in the area of screening was 13.6 deaths/1000 subject years. In the formal comparison of the two 'populations', the true proportions of number of deaths divided by number of subjects (from which the mortality rates are derived) were used. Deaths for these two populations were compared as two binomial proportions, using the formula $(p_1 - p_2) \pm 1.96\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{36n_2}}$, where p_1 was the relative frequency of mortality in the whole of Scotland and p_2 was the relative frequency of mortality in the screening area over the 6 year period as defined above. This led to a 95% confidence interval of (-2.6, -1.9) deaths per 1000 years for the difference between the two

proportions. Thus we could reject the hypothesis that the underlying mortality rate was the same for the two populations.

Similarly, for coronary heart disease, the Scottish national mortality rate was 4.2/1000 years and the screening area population rate was 4.9/1000 subject years, with the confidence interval for the difference being (-1.0, -0.6) deaths per 1000 subject years, and for cancer deaths the national rate was 3.8/1000 subject years and the screening area rate was 4.2/1000 subject years, with a confidence interval of (-0.75, -0.35) deaths per 1000 years for the difference between them. The assumption that the population mortality rate in the screening area was the same as that in Scotland as a whole was thus rejected for these causes of death also. It can be concluded from this that the population mortality rates in the screening area are significantly higher than the rates in Scotland as a whole.

The above formula for a confidence interval for the difference between two proportions assumes, among other things, that the two proportions being compared relate to independent samples. This assumption is not valid since the screening area population is contained within the Scottish national population, and thus these confidence intervals are only approximate. The Scottish national population is around three times the size of the screening area population.

7.3.2 Comparisons within the screening area population

The Carstairs deprivation score (as discussed in section 6.2.1) was assigned to records on the basis of postcode sector. These deprivation scores are based on the 1991 census data. The deprivation scores were grouped into seven categories, known as DEPCATs, as suggested by Carstairs and Morris (1991). The DEPCATs were used to divide the screening area into subpopulations to be compared. To make the seven subpopulations more comparable, the mortality rates were age standardised against the total screening area population. Both the crude and age-standardised mortality rates for each of the

DEPCATs are given in Table 7.2. This analysis was restricted to men age 45-64. The crude mortality rates were calculated as in the previous section. The age-standardised mortality rates were calculated in an analogous fashion based on the age distribution in the standard population. The standard population was taken to be all men age 45-64 in the screening area (based on the 1991 census), and the study populations were taken to be the DEPCAT groups which together made up the standard population.

DEPCAT	Crude mortality rate	Age-standardised mortality rate
1	6.30	6.83
2	7.44	7.72
3	8.56	8.61
4	10.39	10.54
5	11.58	11.50
6	13.45	13.14
7	16.87	16.50

Table 7.2 Crude and age-standardised population mortality rates (events per 1000 years) for each of the DEPCAT areas within the screening area. Standardisation is against the total population in the screening area.

The age-standardised mortality rates were similar to the crude mortality rates in each of the DEPCATs. It can be seen that the more affluent areas (low DEPCATs) had lower mortality rates than the deprived areas (high DEPCATs), as has been found in other studies (McLoone and Boddy, 1994).

7.4 Mortality trends in the screened cohort

Trends in screenee death rates were examined by looking at the mortality rate in each 3 month block of follow-up. Until the end of 1993, the screened subjects were followed

up for between 33 and 63 months, depending on when they were screened. Life tables were constructed for all-cause mortality in the 3 month blocks of follow-up for males, and for females, in each of six age categories - <45 / 45-49 / 50-54 / 55-59 / 60-64 / >64.

The proportion dying during the 3 month block, $p_t = \frac{d_t}{n_t}$ where d_t is the number of deaths occurring in block t and n_t is the number of subjects at risk at the start of the 3 month block of follow-up.

The proportion surviving, $S_t = 1\text{-}p_t$ and the survivor function, $\hat{S}(t) = \prod_{i=1}^{t} S_i$.

It should be noted that the only screenee death information released by the Scottish Record Linkage System (SRL) was a cross-tabulation of 3 month blocks of 'number of days from screening till 1.1.94' and 'number of days from screening till death', thus preserving subject anonymity as discussed in chapter 6.

Plots of the survivor function $\hat{S}(t)$, such as Figure 7.1, showed, as expected, that survival times were larger for females than for males in each age group.

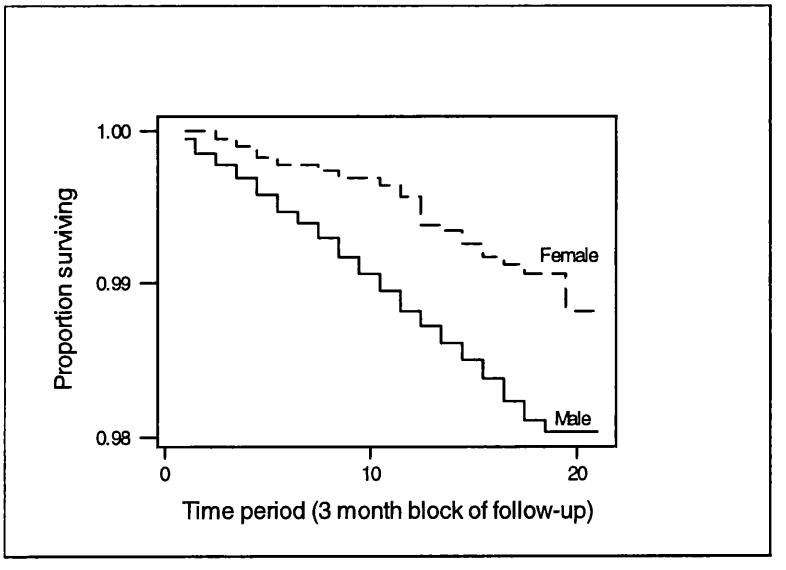


Figure 7.1 Survivor function for 45-49 year old male and female screenees

A similar plot of the survivor function for males, Figure 7.2, showed that survival times were longer for the younger age groups, as expected. The plot for females revealed a similar pattern.

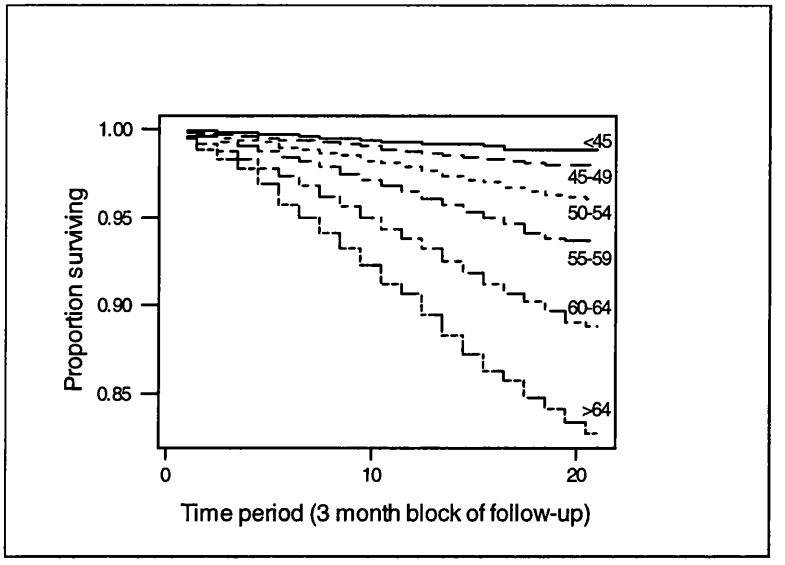


Figure 7.2 Survivor function for male screenees by age group

7.5 Comparison of the screening area population with the screened cohort

The main objective of this chapter is the comparison of mortality rates in the screened cohort with mortality rates in the population in the area of screening. To carry out this comparison, certain assumptions have been made. These include a constant population size (since we have used the population size from the 1991 census for each of the six years from 1988 to 1993), and a lack of any trends in the population or screened cohort mortality rates during the time period considered. The analysis will focus on men aged 45-64.

Initially, consider the first year of subject follow-up for each of the screened men age 45-64. This avoids the problems of an ageing cohort. For the first year, the age group specified for the cohort will be directly comparable to the age group specified for the population. In later years, the cohort will be older, and some subjects will have moved into an older age group. However, screened subjects can only be identified by their age group at screening visit 1, so no adjustment can be made for this ageing. Screening took place between 1988 and 1991. For this reason, the death rate in 1991 was used as the population rate, since it was close to the mid-point of follow-up, and the census data were collected in 1991.

For the first year of follow-up, 95% confidence intervals were produced for all-cause mortality for the difference between the cohort and population mortality rates, in each of the five year age-bands - 45-49 / 50-54 / 55-59 / 60-64. Ages <45 or >64 years were not considered for the calculation of confidence intervals due to the difficulty in defining subgroups which would be directly comparable in the screened cohort. The results were as in Table 7.3.

Age group	95% CI for males	95% CI for females
45-49	-3.52 to -1.52	-3.82 to -1.23
50-54	-5.01 to -2.36	-5.50 to -3.62
55-59	-7.30 to -3.95	-8.75 to -5.06
60-64	-10.16 to -5.49	-11.95 to -5.04

Table 7.3 95% confidence intervals for the difference in screenee and population all-cause mortality rates (events/1000 subject years) for 45-64 year old males and females.

None of the confidence intervals in Table 7.3 contains zero, with the population rate being significantly larger than the screenee rate for men and women age 45-64. This supports the hypothesis that it is more healthy subjects who are willing and or able to come for screening.

However, in later years of follow-up, the population mortality rate may not be significantly higher than the mortality rate in the screened cohort, since subjects who are healthy and willing to come for screening one year may not necessarily still be healthy a year later.

The year 2 mortality rate was calculated as the number of deaths in year 2 divided by the number of subjects at risk at the start of year 2. This approach was also used to calculate the screenee all-cause mortality rates for year 3. The rates calculated by this method are given in Table 7.4.

Sex	Age	Year 1 death rates	Year 2 death rates	Year 3 death rates
Male	<45	1.6	2.8	2.6
	45-49	2.98	3.98	4.89
	50-54	5.53	7.14	8.46
	55-59	8.91	12.35	14.08
	60-64	16.72	21.35	24.92
	>64	21.19	37.51	37.31
Female	<45	0	0.58	1.17
	45-49	0.868	1.74	1.74
	50-54	0.378	2.27	2.65
	55-59	1.92	5.38	5.02
	60-64	6.34	5.01	9.62
	>64	9.15	15.0	18.75

Table 7.4 Screenee mortality rates in each year of follow-up broken down by sex and age group at screening. The death rates are the number of deaths per 1000 subject years.

While the screenee mortality rates do in general increase across the years of follow-up (for the first 3 years at least), it must be remembered that the cohort is also ageing. The group of subjects who were age 45-49 in the first year of follow-up were age 46-50 in the second year of follow-up. In Figure 7.3, a five year age band is taken as reflecting an interval represented by one unit. The mortality rates are plotted against the midpoint of the interval, and the increasing age of the cohort is adjusted for by moving along the axis by 0.2 of an interval (representing 1 year) for each of the years of follow-up. This gave a plot for males as in Figure 7.3, with females showing a similar pattern although the actual rates were lower.

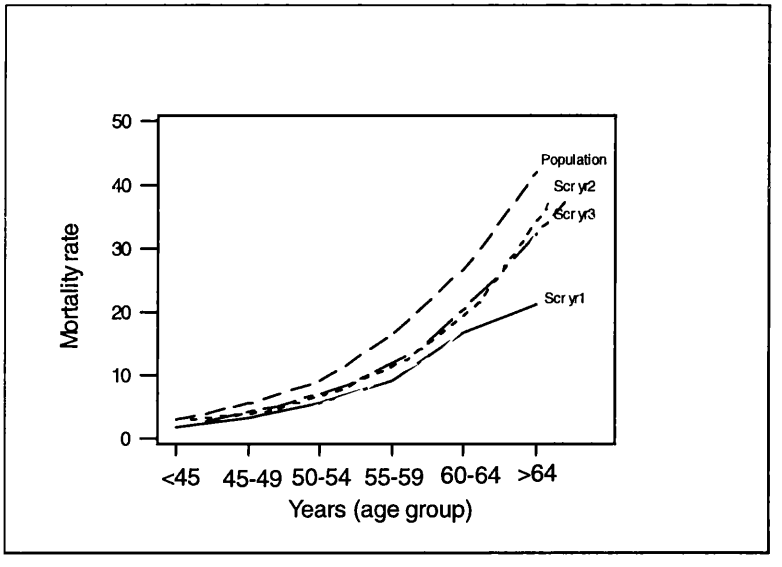


Figure 7.3 Male population and screenee mortality rates (events per 1000 subject years) against age group

It can be seen that the 'shifted' rates are similar for years 2 and 3, while year 1 is slightly lower. The population rate was clearly higher than the screenee rates, even after 3 years of follow-up. While it is not surprising that the screened cohort was more healthy than the population in the first year or two after screening, it is interesting that the rate does not appear to have increased to the population level 3 years after screening.

Since the screening process concentrated on males in the 45-64 year old age group, there was a much smaller group of men aged <45 or >64. These age groups were also defined

differently for the screened cohort and the population. For the screened cohort, any man outwith the 45-64 year age group was included in one of these categories, but for the population these bands were restricted to age 40-44 and age 65-69. These two age bands were thus excluded, and analysis focused on the 45-64 year old group.

When a logistic transform ($\log_e \frac{p}{1-p}$) is applied to each of the all-cause mortality rates, the trends become reasonably linear for males, as can be seen in Figure 7.4. For females the trends with age are more erratic, possibly due to the smaller sample size of women screened.

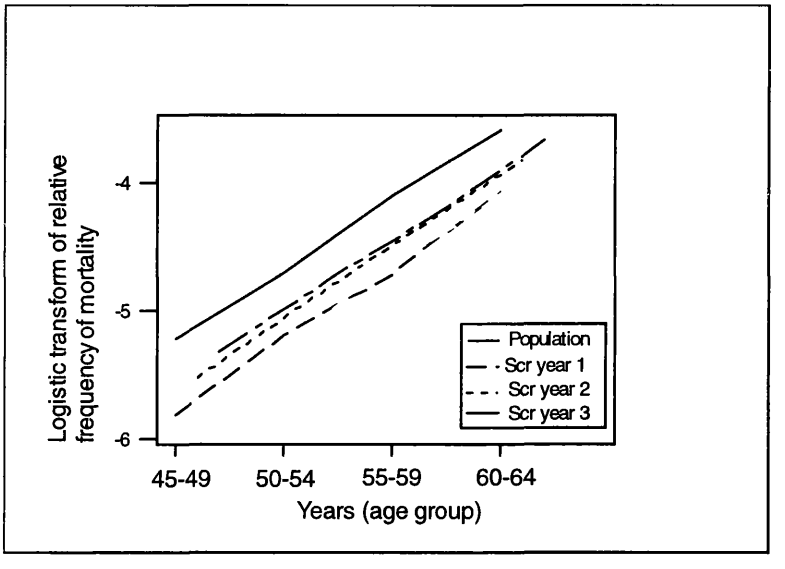


Figure 7.4 Logistic transform of the relative frequencies of all-cause mortality for males in the population and screened cohort against age group.

Logistic regression was used to model the probability of death given age and group. Group 1 related to the screening area population mortality rates for 1991 since the census data were collected in that year, group 2 to the cohort of subjects available for the first year of follow-up, group 3 to the cohort of subjects available for the second year of follow-up, and group 4 to the cohort of subjects available for a third year of follow-up. The term fitted for the age*group interaction was not significant (p=0.1520), so there was no evidence to reject the assumption that each of the groups showed the same

trend with age. Fitting a main effects logistic model showed that age and group were both highly significant ($p<0.001$ for both). Estimated coefficients for the contrasts among the probabilities of death for the different groups are given in Table 7.5. The estimated coefficients divided by standard errors can be considered as z-statistics. It can be seen that the mortality rates in each year of follow-up for the screened cohort were significantly lower than the mortality rates in the screening area population in 1991. Consideration of the population deaths in other years (1990 and 1992) made little difference to this result.

Comparison	Coefficient	Standard error of coefficient	Coeff / standard error
Screenee year 1 - Population	-0.464	0.044	-10.4
Screenee year 2 - Population	-0.295	0.040	-7.33
Screenee year 3 - Population	-0.245	0.038	-6.38

Table 7.5 Coefficients, standard errors and z-statistics (coefficient divided by standard error) for the comparison of screenees in each year of follow-up with the population in the screening area.

7.6 Discussion

The main difficulty in comparing the WOSCOPS screened cohort with the general population in the area of screening has been the need to preserve subject anonymity in the screened cohort. This has meant that subjects were identifiable only by sex and age group at screening. As length of follow-up increased, the cohort was ageing and thus subjects would have moved into a higher age group and the youngest age group would have progressively decreased in size. However, with no knowledge of the age distribution of subjects who died within each five year age group, it was impossible to accurately adjust for this source of bias. This problem has been partially addressed by consideration of 'shifted' mortality rates in the logistic models. It was interesting to note

that the mortality rates in the screened cohort were still lower than those in the general population in the third year of follow-up. It had been expected that in the first year of follow-up, the mortality rates would have been lower due to healthier people attending for screening. However, the mortality rates in the screened cohort did not become closer to those in the general population in the later years of follow-up. This may be a reflection of the fact that subjects who volunteer for public health screening are more health conscious, and the effects of this persist for a long period of time. The length of follow-up considered here may be too short to overcome the 'healthy screenee' effect.

A further source of bias is the fact that groups being compared are not independent since the screened cohort is contained within the population in the screening area. It is hoped that this will not be too major a source of error since the screening area population is larger than the screened cohort (the cohort makes up approximately 35% of the population of men age 45-64 in the screening area). However, this correlation would, if anything, make the groups more alike.

Overall, the screened cohort does not appear to be very representative of the population from which it came, at least for a short period of follow-up, and thus there does seem to be a particular group of healthier people who attend population screening initiatives.

Chapter 8

Summary and future work

8.1 Summary

One of the main findings of this thesis has been the feasibility and accuracy of using record linkage techniques as a means of subject follow-up in clinical trials. The comparative study (dealt with in Chapter 5) for the subjects randomised into the clinical trial branch of WOSCOPS found that record linkage for deaths achieved very high accuracy compared to the individual subject follow-up routinely employed in WOSCOPS. The only death which was missed by record linkage was one which occurred after the patient had moved out of Scotland. While the linkage quality was lower for the incidence databases (hospitalisations and registrations), it was nonetheless still of a reasonably high standard. For example, assuming that all non-psychiatric hospitalisations were identified by one or other system, record linkage identified 94% of events while WOSCOPS individual follow-up identified 91%. Record linkage thus appears to be more successful at identifying adverse events than individual subject follow-up. However, events identified by record linkage should be validated since probabilistic linkage may identify events which do not really belong to the WOSCOPS subjects. Comparisons of the events identified by record linkage using fully postcoded and updated data against unpostcoded data and data recorded at an initial screening visit have found that while linkage quality is reasonable for each data file, it is of a higher standard when updated, postcoded data is used. While the difference in linkage quality with no postcodes and screening visit 1 records is not large at this stage, this difference will increase as the length of subject follow-up increases.

Having established that linkage based on data collected at screening visit 1 gave a reasonable level of accuracy, linkage was used to identify adverse events for all subjects

who attended an initial screening visit. Record linkage provided the only feasible method by which such a large cohort (97,165 subjects with identifying information available) could be followed-up. While analysis was limited by the need to maintain subject anonymity, quite extensive analysis was carried out using categories of the baseline risk factors (see Chapter 6).

The first group studied was the 74,576 men aged 45-64 at screening visit 1- the group targeted by the screening exercise. Analysis by a cross-classification of coronary heart disease risk factors identified the U-shaped relationship between cholesterol and all-cause mortality which has been found in other studies (see Chapter 1), with high cholesterol being associated with risk of coronary heart disease and low cholesterol being associated with risk of cancer mortality. A further point of interest was the higher mortality risk associated with an alcohol consumption of zero units per week. This relationship was also in agreement with other studies. However, the observed relationships become more difficult to interpret once interaction terms are included in the model. These were discussed in more detail in Chapter 6. The relationships observed among risk factors for mortality in these middle-aged men also appeared when hospitalisation was the outcome considered.

It was interesting to note the differences in the results for cancer outcomes when each of the three possible databases (cancer mortality, registration and hospitalisation) were considered. It was reassuring to note that the exclusion of competing causes of death had little effect on the results. The exclusion of the early years of follow-up also had little effect on the results, with the exception of cholesterol, for which the excess risk in the lowest cholesterol quintile decreased as years were excluded. Both these analyses support the suggestion that the relationship between cancer and low cholesterol may, at least in part, be a short term artefact of undiagnosed disease. For fuller discussion of this conclusion see sections 6.3.1.4 and 6.3.1.5.

The second group studied was the cohort of 14,950 women who attended screening visit 1. Consideration of female mortality resulted in trends similar to the analysis of middle-aged men, although the factors fitted were often not statistically significant due to the low event rate in this group of subjects.

The third group studied was the cohort of 13,559 middle-aged men with raised cholesterol who reached screening visit 3. These men were analysed by a cross-classification based on laboratory measurements made at this third screening visit. The main finding of this analysis was that plasma fibrinogen exhibited the same U-shaped curve with mortality as total cholesterol, with the highest quintile being associated with coronary heart disease mortality and the lowest quintile with cancer mortality. However, fibrinogen was not significantly related to coronary heart disease or cancer mortality, once other factors had been taken into account.

Having obtained these results for the screened cohort, it was then of interest to consider how applicable these conclusions are to the general population in the area of screening. From the analysis in Chapter 7, it would appear that the screened cohort of men age 45-64 have lower mortality rates than the population of 45-64 year old men in the screening area. It was expected that it would be more healthy subjects who were able to attend for screening in the first instance, but that this effect would disappear as follow-up continued. However, the screened cohort is still different from the population after three years of follow-up. This persistent difference may be due to subjects who volunteer for public health screening being generally more health-conscious, or it may be a residual effect of the fact that these subjects were well enough to be able to attend for screening in the first instance.

8.2 Future work

While interesting relationships have been established between risk factors such as cholesterol, alcohol and DEPCAT and mortality from cancer and violence as well as cardiovascular causes, it would be of interest to look at relationships between other risk factors and mortality. Other factors of interest could be use of educational and occupational status in place of DEPCAT which is a non subject specific generalisation across each postcode sector. The relationships may change when alternative covariates have been accounted for in the analysis. It may also be of interest to restrict analysis to screened subjects who have no history of coronary heart disease, although this would

reduce the number of factors which could be included in the cross-classification. While some other cross-classifications have been looked at (although not discussed in this thesis), there is clearly huge potential for further analysis.

The main drawback in the analysis of the screened cohort presented in this thesis was the need to preserve subject anonymity. If analysis using the original data, with continuous variables, was possible then this would enhance the power of the analysis to detect relationships. This would be especially helpful in the consideration of screened women, for whom the event rate was much lower than for the men. This may be possible by requesting that an internal analysis be carried out at the Scottish Record Linkage System, so that only the output from the analysis would be released and no data would be given to WOSCOPS researchers. Since there would then be no constraints on the data available for analysis, it would also be possible to carry out survival analysis using the time from screening until event. It may be argued that there would be no ethical problem with this analysis since no release of data would be involved.

The power to identify associations between baseline risk factors and outcome will increase with increasing years of follow-up. This will be particularly important for certain subgroups of the cohort being examined, for example, women, non-smoking males or diabetics, and for less common outcomes such as mortality for specific cancer sites.

For the comparison of mortality rates for the screened cohort with rates for the population in the screening area, it would be interesting to repeat the comparison with further years of follow-up included. This would give an indication of the length of time for which the 'healthy screenee' effect persists. In the public health context, this would also provide an indication of the regularity required in screening exercises if subjects at high risk are to be identified.

8.3 Implications of this thesis

This thesis has provided the first external check on the quality of linkage achieved at the Scottish Record Linkage System for linking patient files to the data held at SRL. This study has shown that a high standard of linkage is possible with only a few subject identifiers (name, sex, date of birth and postcode). This thesis has shown that the linkage quality may be enhanced through the postcoding of all addresses for study subjects and periodic updating of identifying information recorded. These points may be of relevance to any other studies which aim to use computerised record linkage to detect events for subject follow-up.

This thesis has also given indications of interesting relationships between baseline risk factors and mortality which agree with previous studies and should be re-examined once sufficient numbers of events have been accumulated. The population comparisons should also be re-examined to see how many years of follow-up are required before the event rate in the screened cohort becomes the same as the event rate in the population in the screening area.

APPENDICES

A: WOSCOPS subject identification form

SC1
Form Type [0][1]
Visit No. [0][0]
Date of Visit [ ][ ][ ][ ][ ][ ] D D M M Y Y

Patient Identification Form
UD1190

Random No. [ ][ ][ ] • [ ][ ][ ][ ]
Patient No. [ ][ ][ ] • [ ][ ][ ][ ]
Patient Initials _____

SECTION 1 Patient and G.P. Information

1. Patient Surname _____ Sex [1] M [2] F
Forename _____
Date of Birth [ ][ ][ ][ ][ ][ ] D D M M Y Y NHS No. [ ][ ][ ][ ][ ][ ][ ][ ][ ][ ][ ][ ][ ][ ]
Address _____
_____
_____
Patient Telephone ____ — _____

2. G.P. Dr. _____
Centre _____
Address _____

SECTION 2 Additional Flagging Details

Patient Surname at Birth _____
Place of Birth _____
_____
_____
Maiden Name _____ (if applicable)
Address on 29th SEPTEMBER 1939
_____
_____
_____

SECTION 3 Dietary Counselling

Has the patient been given dietary counselling? [1] Yes [2] No

SECTION 4 Lipoprotein Analysis

Has the patient fasted overnight? [1] Yes [2] No

Investigator's Initials _____ Name of Person Completing Form _____ Date of Completion _____

B: Examples of WOSCOPS adverse event forms

Serial No. ☐☐☐☐☐
Form Type 7 1
Visit No. ☐☐
Date of Visit ☐☐☐☐☐☐ D D M M Y Y

Adverse Event Summary
UD1 290

Random No. ☐☐☐•☐☐☐☐
Patient No. ☐☐☐•☐☐☐☐
Patient Initials _____

SECTION 1 Adverse Events NB. If more space is required in any section please attach an OC13

Summarise illness or event

Relevant Past Medical History

Relevant Laboratory Date: Date Test Result Normal Range N/A

Concomitant Medications & Dates

Clinical Impression _____

In your estimation is the event related to trial medication? Yes ☐ No ☐
(Please tick the appropriate box.)

Answer all following questions by circling as appropriate		Yes	No
(1)	Did patient die?	1	2
(2)	Was adverse event life threatening?	1	2
(3)	Did event require or prolong in-patient hospitalisation?	1	2
(4)	Was event permanently disabling?	1	2
(5)	Was event cancer, or overdose?	1	2
(6)	Was therapy permanently withdrawn?	1	2
(7)	Was this event a Cerebro-Vascular Accident/TIA?	1	2

(If answer to question 7 is yes, complete form OC14)

SECTION 2 TO BE COMPLETED BY THE ADVERSE EVENTS COMMITTEE

Primary Diagnosis _____
Secondary Diagnosis _____

ICD CODE(S)	SEVERITY			RELATION TO TRIAL DRUG				TRIAL DRUG STATUS		
	Mild	Moderate	Severe	Related	Possible	Unknown	Unrelated*	Temp Withdrawn	Perm. Withdrawn	Not Withdrawn
☐☐☐☐☐☐	1	2	3	1	2	3	4	1	2	3
☐☐☐☐☐☐	1	2	3	1	2	3	4	1	2	3

Has this event been referred to the End-Points Committee? Yes 1 No 2

Serial No. ☐☐☐☐☐
Form Type 5 2
Visit No. ☐☐
Date of Report ☐☐☐☐☐☐ D D M M Y Y

Hospitalisation Report

UD1-291

Random No. ☐☐☐ • ☐☐☐☐
Patient No. ☐☐☐ • ☐☐☐☐
Patient Initials _____

Was the patient hospitalised for this event? Yes [1] No [2]
(If 'YES' complete Section 1)

SECTION 1 To be completed by Trial Physician

1. Date of admission to hospital ☐☐☐☐☐☐ D D M M Y Y
2. Date of discharge from hospital ☐☐☐☐☐☐ D D M M Y Y
3. Hospital _____
   Address _____
   Consultant _____
4. In your clinical opinion what is the main reason for hospitalisation?
   _____
   Comments _____
   _____
   _____
   _____
5. Did the main reason/symptoms of this condition exist prior to randomisation? Yes [1] No [2] Unknown [3]

SECTION 2 To be completed by Trial Monitor

	CATEGORY	CODES
1. Main reason for hospitalisation _____	☐☐	☐☐☐☐☐☐
2. Other illnesses or events during hospitalisation		
_____	☐☐	☐☐☐☐☐☐
_____	☐☐	☐☐☐☐☐☐
_____	☐☐	☐☐☐☐☐☐
3. Treatment or surgical procedure during hospitalisation		
_____	☐☐	☐☐☐☐☐☐
_____	☐☐	☐☐☐☐☐☐
_____	☐☐	☐☐☐☐☐☐

4. Is the event thought to be cardiac related? Yes [1] No [2]
   If 'Yes' ECGs should be attached

OC19

Serial No. ☐☐☐☐☐

Form Type 8 1

Visit No. ☐☐

Date of Visit ☐☐☐☐☐☐ (D D M M Y Y)

Cancer Form

UD1-591

Random No. ☐☐☐•☐☐☐☐

Patient No. ☐☐☐•☐☐☐☐

Patient Initials _____

SECTION 1

1. **Date treatment commenced** ☐☐☐☐☐☐ (D D M M Y Y)

2. **Site of Tumour** _____ ICD Code ☐☐☐☐

3. **Histology** ICD Code ☐☐☐☐

(Give full details of biopsy report)

_____

_____

_____

4. (a) Did the patient have any illness or symptoms prior to randomisation of relevance to this event?

Yes [1] No [2] Unknown [3]

(b) If 'Yes', please give details _____

_____

_____

5. Is this form accompanied by a Cancer Registration Abstract Card? Yes [1] No [2]

_____

SECTION 2

To be completed by Adverse Events Committee

Clinical Diagnosis _____

ICD Code ☐☐☐☐☐☐

_____ Signature for Adverse Events Committee

_____ Name of Person Completing Form

_____ Date

Serial No. ☐☐☐☐☐
Form Type 4 8
Visit No. ☐☐
Date of Report ☐☐☐☐☐☐ D D M M Y Y

Angioplasty Form

UD1-691

Random No. ☐☐☐•☐☐☐☐
Patient No. ☐☐☐•☐☐☐☐
Patient Initials _____

If not already completed for this event, please prepare Form SQ1.

SECTION 1 Coronary Angioplasty

1. Enter date of PTCA ☐☐☐☐☐☐ D D M M Y Y

2. Location
(Record hospital where procedure was performed.)
[1] GRI [2] WIG [3] Other — (Specify)_____
Consultant _____

3. Vessel dilation

(Please circle all responses as appropriate) Vessels	Attempted Yes	Attempted No	Successful Yes	Successful No
Left mainstream	1	2	1	2
Left anterior descending	1	2	1	2
Left circumflex	1	2	1	2
Right coronary artery	1	2	1	2
Other	1	2	1	2
(Specify)	_____			

4. Record outcome

(Please circle all responses as appropriate)	Yes	No		Yes	No
Successful reperfusion	1	2	Repeat PTCA	1	2
MI	1	2	Unstable angina	1	2
Emergency CABG	1	2	Death	1	2
Elective	1	2	Other (specify) _____		

Investigator's Initials _____ Name of Person Completing Form _____ Date of Completion _____

C: Scottish Morbidity Record forms

SMR1:

Scotland - Inpatient and Day Case Records Summary Sheet — SMR1 Medical In Confidence

Hospital

Time of admission/transfer — Ward (WARD) 157-159

Previously attended this hospital? YES/NO

If YES state year

Address (ADDR) 189-210

Tel. no. — Postcode

Religion

Next of kin (Relationship)

Name

Address

Tel. no. — Postcode

Family Doctor

Name

Address

Tel. no. — Postcode

Provisional Diagnosis on Admission

Hospital Code (HOSP) 1-5

Hospital Case Reference No. (CRN) 6-15

Surname (SN) 16-35

First Forename (FN1) 36-50

Second Forename (FN2) 51-65

Maiden Name (MN) 66-85

Alternative Case Reference No. (ACRN) 86-95

Age ..... Date of Birth (DOB) 96-103

Sex (SEX) 104 Marital State (MRST) 105

Postcode (PC) 106-112

GP Practice Code (GPPC) 113-117

GP GMC Number (GPC) 118-124

Admitted/transferred from (ADTF) 125 Type of Admission (TADM) 126

Date placed on Waiting List (DWL) 127-132

Date of Admission (DOA) 133-138

Date of Discharge (DOD) 139-144

Time of Discharge Discharge Code (DISC) 145

Category of Patient (CAT) 146 Type of Facility (TOF) 147

Specialty (SPEC) 148-151

Consultant Surname (CONS) 152-171

Consultant Initials 172-173 Consultant Code (CONC) 174-180

To be completed by doctor on discharge of patient

(1)	Main Condition (DG1)	(DG1C)	1-66
(2)	Other Conditions (DG2)	(DG2C)	67-132
(3)	(DG3)	(DG3C)	133-198
(4)	(DG4)	(DG4C)	1-66
(5)	(DG5)	(DG5C)	67-132
(6)	(DG6)	(DG6C)	133-198

(1)	Main Operation (OP1)	(OP1C)	1-68
(2)	Other Operations (OP2)	(OP2C)	69-136
(3)	(OP3)	(OP3C)	137-204
(4)	(OP4)	(OP4C)	1-68

Date of Main Operation (DOP1) → Date of Main Operation (DOP1) 69-74

ERROR REPORT COMMENT: Data items specified in boxes below are correct (COM)

Enter field abbreviations

Additional notes: 75-134

Contract No. (CTN) 135-144

Local use (LUSE) 145-150

National use (NUSE) 151-154

SMR4:

IMPORTANT: ONLY TO BE USED FOR INFORMAL PATIENTS
MEDICAL IN CONFIDENCE — MENTAL HEALTH STATISTICS SCOTLAND — SMR4 — REVISED 1 1 86

Part 1 – INPATIENT ADMISSION (to be returned in respect of each inpatient admission or transfer to the care of a Psychiatrist)
On admission complete Part 1 and send top copy to the Information Services Division

593039 16

CARD TYPE [1] 7

HOSPITAL NAME

HOSPITAL CODE 8-12

CASE REFERENCE No. 13-22

DATE OF ADMISSION 23-28

SURNAME 29-40

FORENAME 41

SECOND INITIAL 42

MAIDEN SURNAME 43-54

DATE OF BIRTH 55-62

SEX 63

MARITAL STATE 64

HOME ADDRESS

POSTCODE 65-71

CARD TYPE [2] 7

OCCUPATION 8-10

PREVIOUS PSYCHIATRIC IN-PATIENT CARE (If 'yes' state name of hospital) 11

STATUS ON ADMISSION 1 NOT COMPULSORY 2 HOLIDAY ADMISSION 12

IMMEDIATE SOURCE OF REFERRAL 13-14

RESIDENCE IMMEDIATELY PRIOR TO ADMISSION 15

INJURIES OR POISONING PRECIPITATING ADMISSION (STATE TYPE/CAUSE) 16

DIAGNOSIS ON ADMISSION
MAIN PSYCHIATRIC CONDITION 17-22

OTHER CONDITIONS
1 23-28
2 29-34
35-40 41-46
47-52

The hospital should retain the under copy for its own use.

DATE OF DISCHARGE 29-34

OCCUPATIONAL CONDITION ON DISCHARGE 35

TYPE OF DISCHARGE 36

DISPOSAL ON DISCHARGE 37-38

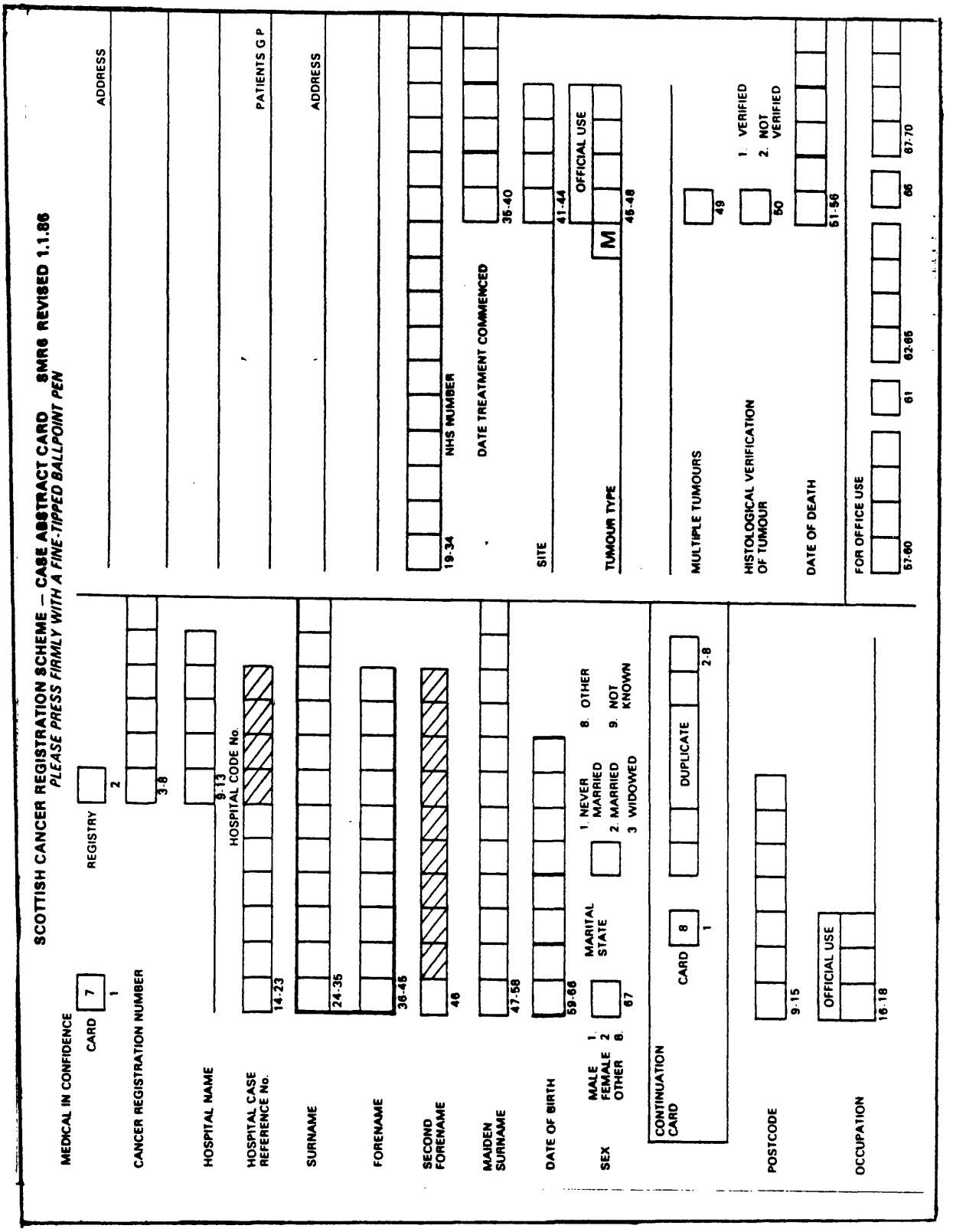
39-44 45-50

ARRANGEMENTS FOR AFTER-CARE - REFERRED TO:
GENERAL PRACTITIONER 51
OUTPATIENT CLINIC 52
LOCAL / REGIONAL AUTHORITY MENTAL HEALTH SERVICES 53
COMMUNITY MENTAL HEALTH SERVICES 54
DAY HOSPITAL UNIT 55
OTHER CARE (SPECIFY) 56
DISCHARGED TO STATUTORY GUARDIANSHIP 57

CARD TYPE [4]

DIAGNOSIS ON DISCHARGE
MAIN PSYCHIATRIC CONDITION 8-13
OTHER CONDITIONS 1 14-19
2 20-25
3 26-31
IF DEAD, REGISTERED PRIMARY CAUSE OF DEATH

SMR6:


SCOTTISH CANCER REGISTRATION SCHEME — CASE ABSTRACT CARD SMR6 REVISED 1.1.86
PLEASE PRESS FIRMLY WITH A FINE-TIPPED BALLPOINT PEN
MEDICAL IN CONFIDENCE
CARD 7 1
REGISTRY 2
CANCER REGISTRATION NUMBER 3-8
HOSPITAL NAME 9-13
HOSPITAL CODE No.
HOSPITAL CASE REFERENCE No. 14-23
SURNAME 24-35
FORENAME 36-45
SECOND FORENAME 46
MAIDEN SURNAME 47-58
DATE OF BIRTH 59-66
SEX MALE 1 FEMALE 2 OTHER 8
MARITAL STATE 67
1. NEVER MARRIED 2. MARRIED 3 WIDOWED 8. OTHER 9. NOT KNOWN
CONTINUATION CARD
CARD 8 1
DUPLICATE 2-8
POSTCODE 9-15
OCCUPATION
OFFICIAL USE 16-18
ADDRESS
PATIENTS G P
ADDRESS
NHS NUMBER 19-34
DATE TREATMENT COMMENCED 35-40
SITE 41-44
OFFICIAL USE
TUMOUR TYPE M 45-48
MULTIPLE TUMOURS 49
HISTOLOGICAL VERIFICATION OF TUMOUR 50
1. VERIFIED 2. NOT VERIFIED
DATE OF DEATH 51-56
FOR OFFICE USE 57-60 61 62-65 66 67-70

SMR20:

PLEASE TYPE OR PRESS FIRMLY WITH A FINE TIPPED BALL POINT PEN. COPY 1 SMR20

MEDICAL IN CONFIDENCE Form Serial No.

Scottish Cardiac Surgery Register: Registration Form

A General Information

1 SCSR Number

2 Surname

3 First Name

4 Other Initial 5 Sex (1 = Male, 2 = Female, 8 = Other, 9 = Not Known)

6 Date of Birth

7 NHS Number*

* Maiden Name
(for married women with NO NHS number)

8 Home Address

9 Postcode

B Administrative Data

14 Accepting Hospital Code

15 Accepting Case Ref. No.

16 Date Accepted for Surgery
(leave blank if admitted as emergency)

20 Urgency Code (1 = Emergency, 2 = Urgent, 3 = Routine, 4 = Deferred)

21 Revised Urgency Code

22 Date of Revision

C Clinical Data

Cardiac Diagnosis

23(i)

24(ii)

25(iii)

Procedure(s) required

27(I)

28(ii)

D Exit from Waiting List

29 Reason

0 = Admitted as Emergency - not on Waiting List
1 = Admitted as Emergency from Waiting List
2 = Admitted from Waiting List
3 = Removed - Transfer to other waiting list - State:

4 = Died 5 = Refused Admission
6 = Removed - Clinical Reasons 7 = Lost Trace 8 = Other - State:

30 Date of Exit from Waiting List

E Details of Stay in Hospital

31 Name of Consultant Surgeon

32 Date of First Operation

Procedure(s) Performed and Procedure Category
1 = Closed 2 = Open Simple
3 = Open Intermediate 4 = Open Complex

33 (i)

34 (ii)

Additional Procedures Performed

35 (i)

Date of this procedure

36 (ii)

Date of this procedure

37 (iii)

Date of this procedure

38 Date of Discharge from Cardiac Surgery Unit

F Disposal from Cardiac Surgery Unit

41 Destination
1 = Home 3 = Other Hospital
2 = Other Unit this Hospital 6 = Died 8 = Other

G Follow Up

44 Date of Death

H Option Boxes

45 National

46 Local

PLEASE NOTE: COMPLETION OF SHADED BOXES IS OPTIONAL

Revised 1993

D: Linkage algorithm

The details of the linkage algorithm used in the SRL are as follows:

1. Locate Soundex code and Soundex weight for each surname in a look-up table and append to the WOSCOPS record.
2. Read the WOSCOPS patient identifying information into an array.
3. Block the array on Soundex code and first initial, and date of birth, storing the position in the WOSCOPS array at which each Soundex code and first initial, and date of birth occurs.
4. Read in records from the ISD linked database, a patient record set at a time, where a patient record set contains all the records linked together for a given patient.
5. If the Soundex code and first initial, or the date of birth is present in the blocking array, then compare each record in the patient record set with each of the WOSCOPS records in this block, and calculate a comparison score for each pair of records.
6. Store the highest score achieved.
7. If this highest score is above the nominal threshold score (total weight), then link the entire patient record set to the WOSCOPS subject with which this highest score is achieved.
8. If more than one WOSCOPS subject achieves the same high score with a patient record set, then link the patient record set with the first WOSCOPS subject to achieve this highest score.
9. Read in the next patient record set.

The specific details of the bonuses and deductions used for the WOSCOPS linkages are as follows:

1. Initialise the matching score to 0.
2. If the Soundex code agrees, add the Soundex weight to the score, where the Soundex weights reflect the rarity of the surname, else make a deduction of 3.0.

3. If the first initial agrees then add a bonus to the score according to the look-up array of first initial bonuses, else make a deduction of 4.9.

4. If the first 4 letters of the forename agree, add a bonus of 3.0 to the score, else make a deduction of 2.5 [for death, SMR6 and SMR20 records where full forename is available].

5. If the sex does not agree, make a deduction of 6.5.

6. If the first 8 characters of the surname do not agree, make a deduction of 2.0.

7. If the year of birth agrees, add a bonus of 6.3. Else make a deduction according to the size of the difference (from a look-up array) up to 20 years, at which the deduction is 7.0.

8. If the month of birth agrees, add a bonus of 3.56, else make a deduction according to the size of the difference (from a look-up array).

9. If the day of birth agrees, add a bonus of 4.9, else make a deduction according to the size of the difference (from a look-up array).

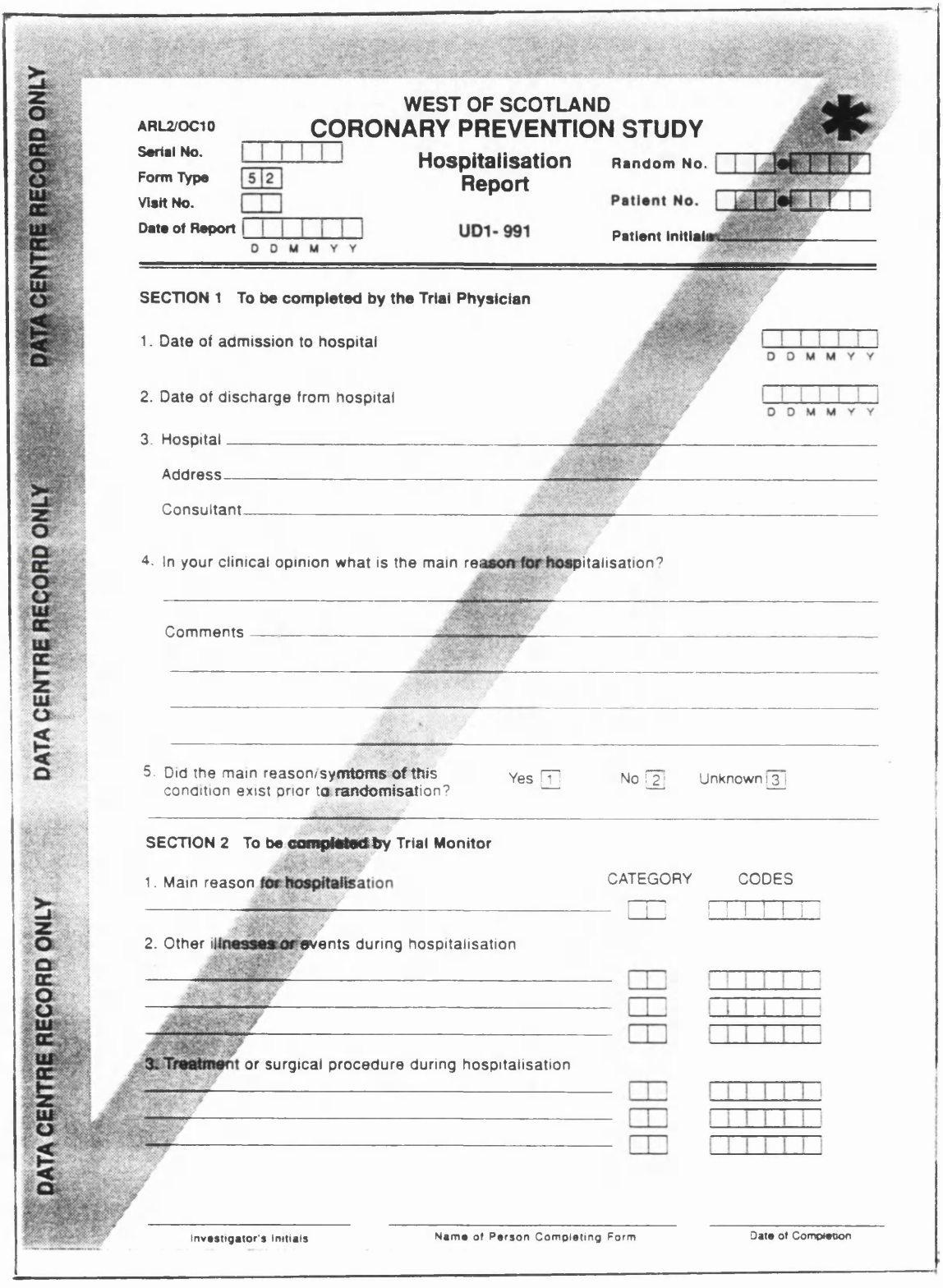
10. Compare the postcode. Add different weights depending on how many characters of the postcode agree and how common each area is (for example, EH gets a larger bonus than G, reflecting the fact that a Glasgow postcode is more common than an Edinburgh one, for the Scottish population as a whole). However, make no deduction for differing postcodes.

11. If the score is between 15 and 35, compare the full surname character by character, and make a further deduction of 2.5 if there are two differences in the name.

12. Add and subtract these individual weights to give the Total Weight on the basis of which a linkage decision is made.

E: Example of a WOSCOPS adverse event form for an event identified only by record linkage

These forms were identical to the standard WOSCOPS adverse event forms, except that they were marked with blue stripes to indicate that they were for Data Centre use only.



DATA CENTRE RECORD ONLY

DATA CENTRE RECORD ONLY

DATA CENTRE RECORD ONLY

ARL2/OC10

WEST OF SCOTLAND
CORONARY PREVENTION STUDY

Serial No.

Form Type 5 2

Visit No.

Date of Report D D M M Y Y

Hospitalisation Report

UD1- 991

Random No.

Patient No.

Patient Initials

SECTION 1 To be completed by the Trial Physician

1. Date of admission to hospital D D M M Y Y

2. Date of discharge from hospital D D M M Y Y

3. Hospital

Address

Consultant

4. In your clinical opinion what is the main reason for hospitalisation?

Comments

5. Did the main reason/symptoms of this condition exist prior to randomisation? Yes 1 No 2 Unknown 3

SECTION 2 To be completed by Trial Monitor

1. Main reason for hospitalisation CATEGORY CODES

2. Other illnesses or events during hospitalisation

3. Treatment or surgical procedure during hospitalisation

Investigator's Initials

Name of Person Completing Form

Date of Completion

F: Covariance matrices

The covariance matrices produced when the final models are fitted may be used in conjunction with the parameter estimates, given in Chapter 6, to examine any possible contrast between levels of the factors. The 'final models' are those arrived at via stepwise logistic regression. For each of these models, the covariance matrices were as follows.

1. All-cause mortality in the 73,110 men aged 45-64 at screening visit 1

	age	smok(1)	smok(2)	dep(1)	dep(2)	alc(1)	alc(2)	alc(3)	alc(4)
age	0.01335								
smok(1)	0.00814	0.04068							
smok(2)	0.00831	0.01967	0.02796						
dep(1)	-0.00023	0.01102	0.01103	0.01533					
dep(2)	-0.00046	0.01065	0.01066	0.01155	0.01869				
alc(1)	0.00012	-0.00011	-0.00002	0.00007	0.00011	0.00177			
alc(2)	0.00019	-0.00027	-0.00027	0.00006	-0.00000	0.00121	0.00491		
alc(3)	0.00015	-0.00020	-0.00031	-0.00002	-0.00009	0.00121	0.00126	0.00919	
alc(4)	0.00020	-0.00024	-0.00039	0.00002	0.00000	0.00121	0.00126	0.00127	0.01475
alc(5)	0.00018	-0.00018	-0.00038	-0.00002	-0.00012	0.00121	0.00126	0.00129	0.00128
alc(6)	0.00017	-0.00017	-0.00041	0.00000	-0.00016	0.00121	0.00128	0.00131	0.00131
chol(1)	0.00001	-0.00007	-0.00003	-0.00000	0.00003	-0.00003	-0.00001	-0.00002	0.00002
chol(2)	-0.00004	-0.00007	-0.00009	0.00002	0.00005	-0.00005	-0.00001	-0.00003	0.00001
chol(3)	-0.00005	-0.00009	-0.00009	-0.00002	0.00003	-0.00004	-0.00002	-0.00002	0.00002
chol(4)	-0.00005	-0.00013	-0.00016	-0.00004	0.00007	-0.00003	-0.00001	-0.00000	0.00002
bmi	-0.00006	0.00426	0.00426	-0.00055	-0.00097	0.00002	-0.00006	-0.00006	-0.00010
dbp	0.00279	-0.00006	0.00022	-0.00002	-0.00004	-0.00006	-0.00016	-0.00026	-0.00016
B*C(1)	0.00022	-0.02575	-0.01103	-0.01532	-0.01154	0.00002	0.00003	0.00002	0.00005
B*C(2)	0.00023	-0.01102	-0.01773	-0.01533	-0.01154	-0.00006	-0.00003	-0.00008	0.00003
B*C(3)	0.00044	-0.02528	-0.01066	-0.01154	-0.01867	0.00003	0.00008	-0.00000	0.00001
B*C(4)	0.00044	-0.01065	-0.01736	-0.01154	-0.01867	-0.00006	-0.00002	-0.00006	-0.00007
B*F(1)	0.00009	-0.00664	-0.00424	0.00055	0.00097	0.00000	0.00005	0.00001	0.00006
B*F(2)	0.00002	-0.00425	-0.00518	0.00055	0.00096	0.00003	0.00014	0.00016	0.00019
A*G	-0.00380	0.00004	-0.00028	0.00001	0.00003	0.00001	0.00004	0.00018	-0.00003
A*B(1)	-0.01116	-0.01430	-0.00818	0.00022	0.00044	-0.00005	-0.00001	-0.00002	-0.00002
A*B(2)	-0.01137	-0.00817	-0.00994	0.00022	0.00043	-0.00008	-0.00005	0.00001	-0.00003
CONSTANT	-0.00984	-0.01946	-0.01965	-0.01105	-0.01073	-0.00120	-0.00106	-0.00093	-0.00099

	alc(5)	alc(6)	chol(1)	chol(2)	chol(3)	chol(4)	bmi	dbp	B*C(1)
alc(5)	0.02018								
alc(6)	0.00133	0.01525							
chol(1)	-0.00002	0.00008	0.00311						
chol(2)	-0.00002	0.00009	0.00134	0.00303					
chol(3)	0.00002	0.00008	0.00135	0.00136	0.00313				
chol(4)	-0.00001	0.00011	0.00135	0.00136	0.00137	0.00282			
bmi	-0.00013	-0.00004	-0.00006	-0.00008	-0.00009	-0.00014	0.00844		
dbp	-0.00024	-0.00028	-0.00003	-0.00007	-0.00009	-0.00013	-0.00021	0.00489	
B*C(1)	-0.00001	0.00003	0.00006	-0.00000	0.00003	0.00002	0.00055	0.00001	0.03223
B*C(2)	-0.00004	-0.00008	0.00004	0.00005	0.00005	0.00006	0.00054	0.00004	0.01532
B*C(3)	0.00002	-0.00002	0.00005	0.00002	0.00002	0.00003	0.00096	0.00004	0.02651
B*C(4)	-0.00009	-0.00015	0.00006	0.00010	0.00012	0.00010	0.00096	0.00005	0.01154
B*F(1)	0.00006	-0.00001	0.00003	0.00004	0.00004	0.00003	-0.00840	0.00009	-0.00073
B*F(2)	0.00021	0.00015	-0.00001	-0.00004	-0.00008	-0.00008	-0.00840	-0.00002	-0.00055
A*G	0.00014	0.00014	-0.00001	0.00002	0.00003	0.00006	0.00004	-0.00486	0.00002
A*B(1)	-0.00006	-0.00005	-0.00004	0.00001	-0.00002	-0.00001	0.00003	0.00000	-0.00039
A*B(2)	0.00006	0.00015	-0.00002	0.00004	0.00005	0.00006	0.00003	-0.00028	-0.00023
CONSTANT	-0.00092	-0.00099	-0.00128	-0.00124	-0.00119	-0.00114	-0.00407	-0.00256	0.01099

	B*C(2)	B*C(3)	B*C(4)	B*F(1)	B*F(2)	A*G	A*B(1)	A*B(2)	CONSTANT
B*C(2)	0.02313								
B*C(3)	0.01154	0.03718							
B*C(4)	0.01830	0.01867	0.02683						
B*F(1)	-0.00055	-0.00116	-0.00097	0.01308					
B*F(2)	-0.00055	-0.00097	-0.00091	0.00840	0.01078				
A*G	-0.00001	0.00000	-0.00002	-0.00010	0.00004	0.00658			
A*B(1)	-0.00022	-0.00073	-0.00044	-0.00001	-0.00003	0.00004	0.01887		
A*B(2)	-0.00029	-0.00044	-0.00054	-0.00004	0.00001	0.00040	0.01114	0.01376	
CONSTANT	0.01101	0.01059	0.01061	0.00416	0.00425	0.00273	0.00822	0.00837	0.02280

2. CHD mortality in the 73,110 men aged 45-64 at screening visit 1

	age	smok(1)	smok(2)	chol(1)	chol(2)	chol(3)	chol(4)	dep(1)	dep(2)
age	0.03066								
smok(1)	0.02422	0.12935							
smok(2)	0.02423	0.06432	0.09128						
chol(1)	0.00001	-0.00020	-0.00012	0.00963					
chol(2)	-0.00008	-0.00016	-0.00028	0.00484	0.00896				
chol(3)	-0.00010	-0.00017	-0.00042	0.00485	0.00486	0.00887			
chol(4)	-0.00006	-0.00030	-0.00063	0.00485	0.00487	0.00488	0.00768		
dep(1)	-0.00057	0.02893	0.02894	-0.00003	0.00003	-0.00008	-0.00012	0.03826	
dep(2)	-0.00127	0.02856	0.02855	0.00005	0.00013	0.00007	0.00015	0.02909	0.04690
alc(1)	0.00159	0.01609	0.01610	-0.00008	-0.00012	-0.00016	-0.00015	0.00089	0.00158
alc(2)	0.00417	0.01673	0.01675	-0.00012	-0.00011	-0.00018	-0.00032	0.00094	-0.00088
alc(3)	0.00483	0.01479	0.01480	-0.00018	-0.00029	-0.00045	-0.00017	-0.00121	-0.00338
alc(4)	0.00337	0.01511	0.01516	-0.00003	0.00005	-0.00032	-0.00064	0.00026	-0.00097
alc(5)	0.00518	0.01443	0.01447	-0.00008	-0.00004	-0.00038	-0.00055	-0.00150	-0.00398
alc(6)	0.00544	0.01556	0.01555	-0.00018	-0.00007	0.00001	-0.00002	-0.00061	-0.00562
dbp	-0.00113	0.01056	0.01057	-0.00011	-0.00017	-0.00019	-0.00025	-0.00025	-0.00038
A*B(1)	-0.03066	-0.04018	-0.02423	-0.00007	0.00003	-0.00000	0.00002	0.00057	0.00127
A*B(2)	-0.03066	-0.02423	-0.02866	-0.00002	0.00014	0.00021	0.00025	0.00057	0.00127
B*F(1)	0.00113	-0.01616	-0.01055	0.00009	0.00012	0.00002	0.00003	0.00025	0.00038
B*F(2)	0.00113	-0.01055	-0.01357	-0.00001	-0.00003	-0.00004	-0.00004	0.00025	0.00038
B*E(1)	-0.00159	-0.02569	-0.01607	0.00001	-0.00003	-0.00005	-0.00005	-0.00088	-0.00158
B*E(2)	-0.00159	-0.01609	-0.02216	-0.00000	0.00001	0.00012	0.00016	-0.00089	-0.00158
B*E(3)	-0.00417	-0.02694	-0.01673	-0.00001	-0.00011	-0.00002	0.00011	-0.00094	0.00088
B*E(4)	-0.00417	-0.01673	-0.02294	0.00014	0.00017	0.00023	0.00044	-0.00094	0.00089
B*E(5)	-0.00483	-0.02381	-0.01476	-0.00001	0.00011	0.00015	-0.00028	0.00122	0.00337
B*E(6)	-0.00483	-0.01480	-0.02079	0.00018	0.00029	0.00055	0.00035	0.00121	0.00338
B*E(7)	-0.00337	-0.02453	-0.01514	-0.00013	-0.00019	0.00021	0.00037	-0.00026	0.00096
B*E(8)	-0.00338	-0.01512	-0.02111	0.00015	0.00004	0.00050	0.00094	-0.00027	0.00097
B*E(9)	-0.00517	-0.02321	-0.01443	0.00009	-0.00037	0.00025	0.00010	0.00151	0.00396
B*E(10)	-0.00519	-0.01444	-0.02066	0.00002	0.00008	0.00060	0.00076	0.00150	0.00398
B*E(11)	-0.00544	-0.02400	-0.01553	-0.00001	-0.00021	-0.00052	-0.00018	0.00061	0.00562
B*E(12)	-0.00545	-0.01558	-0.02150	0.00056	0.00052	0.00051	0.00055	0.00061	0.00563
B*D(1)	0.00058	-0.07124	-0.02893	0.00015	0.00001	0.00007	0.00005	-0.03826	-0.02909
B*D(2)	0.00057	-0.02893	-0.04650	0.00010	0.00014	0.00013	0.00016	-0.03826	-0.02909
B*D(3)	0.00127	-0.07096	-0.02857	0.00016	0.00009	0.00008	0.00009	-0.02909	-0.04689
B*D(4)	0.00126	-0.02857	-0.04610	0.00016	0.00027	0.00031	0.00027	-0.02910	-0.04689
CONSTANT	-0.02418	-0.06413	-0.06401	-0.00472	-0.00467	-0.00452	-0.00441	-0.02889	-0.02864

	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)	alc(6)	dbp	A*B(1)	A*B(2)
alc(1)	0.02332								
alc(2)	0.01397	0.09256							
alc(3)	0.01392	0.01469	0.26794						
alc(4)	0.01393	0.01449	0.01466	1.02039					
alc(5)	0.01392	0.01481	0.01521	0.01484	0.52480				
alc(6)	0.01386	0.01488	0.01526	0.01471	0.01542	1.02370			
dbp	0.00003	-0.00138	-0.00237	-0.00236	-0.00293	-0.00134	0.02060		
A*B(1)	-0.00159	-0.00417	-0.00483	-0.00337	-0.00518	-0.00544	0.00113	0.05019	
A*B(2)	-0.00160	-0.00418	-0.00483	-0.00340	-0.00520	-0.00544	0.00112	0.03066	0.03722
B*F(1)	-0.00003	0.00140	0.00238	0.00239	0.00295	0.00134	-0.02060	-0.00147	-0.00113

	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)	alc(6)	dbp	A*B(1)	A*B(2)
B*F(2)	-0.00003	0.00140	0.00238	0.00238	0.00295	0.00134	-0.02059	-0.00113	-0.00131
B*E(1)	-0.02331	-0.01396	-0.01391	-0.01391	-0.01390	-0.01386	-0.00002	0.00224	0.00159
B*E(2)	-0.02332	-0.01397	-0.01391	-0.01394	-0.01393	-0.01386	-0.00003	0.00159	0.00184
B*E(3)	-0.01397	-0.09255	-0.01468	-0.01448	-0.01480	-0.01488	0.00139	0.00608	0.00417
B*E(4)	-0.01397	-0.09257	-0.01469	-0.01450	-0.01482	-0.01488	0.00138	0.00417	0.00490
B*E(5)	-0.01391	-0.01467	-0.26793	-0.01461	-0.01517	-0.01526	0.00239	0.00669	0.00482
B*E(6)	-0.01392	-0.01470	-0.26795	-0.01468	-0.01523	-0.01525	0.00237	0.00483	0.00590
B*E(7)	-0.01393	-0.01448	-0.01465	-1.02037	-0.01482	-0.01471	0.00236	0.00493	0.00339
B*E(8)	-0.01394	-0.01450	-0.01466	-1.02043	-0.01487	-0.01471	0.00235	0.00337	0.00404
B*E(9)	-0.01391	-0.01479	-0.01520	-0.01480	-0.52477	-0.01542	0.00295	0.00706	0.00518
B*E(10)	-0.01392	-0.01482	-0.01522	-0.01488	-0.52483	-0.01541	0.00292	0.00518	0.00647
B*E(11)	-0.01385	-0.01488	-0.01523	-0.01470	-0.01540	-1.02370	0.00135	0.00714	0.00543
B*E(12)	-0.01387	-0.01490	-0.01528	-0.01475	-0.01545	-1.02370	0.00132	0.00544	0.00680
B*D(1)	-0.00089	-0.00094	0.00121	-0.00024	0.00152	0.00061	0.00025	-0.00105	-0.00058
B*D(2)	-0.00089	-0.00094	0.00121	-0.00025	0.00151	0.00061	0.00025	-0.00057	-0.00079
B*D(3)	-0.00158	0.00088	0.00337	0.00095	0.00397	0.00562	0.00038	-0.00208	-0.00127
B*D(4)	-0.00159	0.00087	0.00336	0.00094	0.00395	0.00562	0.00037	-0.00126	-0.00163
CONSTANT	-0.01598	-0.01657	-0.01458	-0.01490	-0.01421	-0.01551	-0.01041	0.02423	0.02411

	B*F(1)	B*F(2)	B*E(1)	B*E(2)	B*E(3)	B*E(4)	B*E(5)	B*E(6)	B*E(7)
B*F(1)	0.03192								
B*F(2)	0.02060	0.02640							
B*E(1)	-0.00044	0.00003	0.03689						
B*E(2)	0.00002	-0.00028	0.02331	0.03178					
B*E(3)	-0.00252	-0.00140	0.02249	0.01397	0.13532				
B*E(4)	-0.00140	-0.00200	0.01396	0.01995	0.09256	0.11653			
B*E(5)	-0.00352	-0.00238	0.02239	0.01390	0.02337	0.01466	0.61196		
B*E(6)	-0.00238	-0.00301	0.01391	0.01991	0.01469	0.02081	0.26792	0.30124	
B*E(7)	-0.00385	-0.00238	0.02242	0.01394	0.02318	0.01449	0.02335	0.01467	1.36630
B*E(8)	-0.00239	-0.00324	0.01391	0.01993	0.01448	0.02061	0.01459	0.02082	1.02039
B*E(9)	-0.00397	-0.00295	0.02236	0.01392	0.02347	0.01479	0.02394	0.01521	0.02352
B*E(10)	-0.00296	-0.00358	0.01390	0.01992	0.01480	0.02097	0.01516	0.02145	0.01485
B*E(11)	-0.00283	-0.00134	0.02227	0.01385	0.02361	0.01487	0.02414	0.01522	0.02354
B*E(12)	-0.00133	-0.00218	0.01386	0.01985	0.01489	0.02108	0.01525	0.02154	0.01473
B*D(1)	-0.00026	-0.00025	0.00164	0.00089	0.00179	0.00094	-0.00125	-0.00121	0.00095
B*D(2)	-0.00025	-0.00016	0.00088	0.00095	0.00094	0.00109	-0.00122	-0.00156	0.00025
B*D(3)	-0.00021	-0.00038	0.00281	0.00158	-0.00017	-0.00088	-0.00433	-0.00337	-0.00086
B*D(4)	-0.00038	-0.00027	0.00158	0.00181	-0.00088	-0.00092	-0.00338	-0.00387	-0.00094
CONSTANT	0.01050	0.01058	0.01610	0.01602	0.01672	0.01652	0.01480	0.01452	0.01503

	B*E(8)	B*E(9)	B*E(10)	B*E(11)	B*E(12)	B*D(1)	B*D(2)	B*D(3)	B*D(4)
B*E(8)	1.07354								
B*E(9)	0.01482	0.79013							
B*E(10)	0.02108	0.52479	0.61054						
B*E(11)	0.01469	0.02430	0.01538	1.28840					
B*E(12)	0.02099	0.01543	0.02180	1.02368	1.09079				
B*D(1)	0.00024	-0.00189	-0.00152	-0.00024	-0.00060	0.08527			
B*D(2)	0.00042	-0.00152	-0.00166	-0.00062	-0.00084	0.03826	0.05849		
B*D(3)	-0.00095	-0.00528	-0.00397	-0.00789	-0.00561	0.07135	0.02910	0.09719	
B*D(4)	-0.00117	-0.00397	-0.00461	-0.00564	-0.00670	0.02909	0.04674	0.04689	0.06829
CONSTANT	0.01476	0.01440	0.01412	0.01573	0.01514	0.02887	0.02883	0.02848	0.02837

	CONSTANT
CONSTANT	0.06789

3. Cancer mortality in the 73,110 men aged 45-64 at screening visit 1

	age	smok(1)	smok(2)	bmi	chol(1)	chol(2)	chol(3)	chol(4)	dep(1)
age	0.00469								
smok(1)	-0.00030	0.00957							
smok(2)	0.00002	0.00619	0.00774						
bmi	0.00002	-0.00014	0.00052	0.00376					
chol(1)	-0.00005	-0.00009	-0.00003	-0.00015	0.00761				
chol(2)	-0.00003	-0.00010	-0.00003	-0.00026	0.00308	0.00742			
chol(3)	-0.00003	-0.00017	-0.00004	-0.00037	0.00309	0.00311	0.00834		
chol(4)	0.00001	-0.00029	-0.00020	-0.00049	0.00310	0.00312	0.00314	0.00811	
dep(1)	-0.00013	-0.00041	-0.00063	-0.00009	0.00008	0.00011	0.00004	-0.00000	0.00990
dep(2)	-0.00022	-0.00045	-0.00100	-0.00007	0.00021	0.00034	0.00032	0.00037	0.00838
CONSTANT	-0.00334	-0.00541	-0.00574	-0.00142	-0.00306	-0.00311	-0.00300	-0.00286	-0.00777

	dep(2)	CONSTANT
dep(2)	0.01091	
CONSTANT	-0.00771	0.01791

4. Trauma mortality in the 73,110 men aged 45-64 at screening visit 1

	age	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)	alc(6)	CONSTANT
age	0.03417							
alc(1)	0.00202	0.04595						
alc(2)	0.00479	0.03160	0.15697					
alc(3)	0.00584	0.03166	0.03213	0.19932				
alc(4)	0.00410	0.03156	0.03189	0.03202	0.53222			
alc(5)	0.00640	0.03169	0.03221	0.03241	0.03208	0.20030		
alc(6)	0.00684	0.03172	0.03227	0.03248	0.03214	0.03260	0.36684	
CONSTANT	-0.02336	-0.03270	-0.03459	-0.03531	-0.03412	-0.03569	-0.03599	0.04729

5. Hospitalisation for any cause for the men aged 45-64 at screening visit 1

	age	smok(1)	smok(2)	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)	alc(6)
age	0.00230								
smok(1)	0.00045	0.00103							
smok(2)	0.00041	0.00049	0.00077						
alc(1)	0.00002	-0.00002	-0.00001	0.00038					
alc(2)	0.00003	-0.00004	-0.00004	0.00027	0.00088				
alc(3)	0.00004	-0.00004	-0.00005	0.00027	0.00028	0.00176			
alc(4)	0.00001	-0.00003	-0.00006	0.00027	0.00028	0.00028	0.00319		
alc(5)	0.00004	-0.00003	-0.00006	0.00027	0.00028	0.00029	0.00029	0.00483	
alc(6)	0.00002	-0.00003	-0.00008	0.00027	0.00029	0.00029	0.00029	0.00029	0.00369
dep(1)	0.00085	-0.00004	-0.00007	0.00001	0.00001	-0.00001	-0.00001	-0.00002	-0.00002
dep(2)	0.00083	-0.00004	-0.00012	0.00002	-0.00000	-0.00003	-0.00002	-0.00003	-0.00008
chol(1)	-0.00000	-0.00000	0.00000	-0.00000	-0.00000	-0.00000	-0.00000	-0.00001	0.00001
chol(2)	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	-0.00000	-0.00001	0.00001
chol(3)	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	-0.00000	0.00001
chol(4)	-0.00001	-0.00002	-0.00002	-0.00001	-0.00001	-0.00001	-0.00001	-0.00001	0.00002
A*D(1)	-0.00169	0.00004	0.00007	0.00000	0.00000	0.00001	0.00002	0.00001	0.00002
A*D(2)	-0.00165	0.00004	0.00012	-0.00001	0.00001	0.00001	0.00002	0.00000	0.00005
A*B(1)	-0.00086	-0.00102	-0.00049	-0.00001	-0.00000	-0.00001	-0.00001	-0.00001	-0.00001
A*B(2)	-0.00080	-0.00049	-0.00076	-0.00001	-0.00001	-0.00000	-0.00001	0.00001	0.00002
CONSTANT	-0.00118	-0.00042	-0.00039	-0.00027	-0.00026	-0.00024	-0.00024	-0.00024	-0.00024

	dep(1)	dep(2)	chol(1)	chol(2)	chol(3)	chol(4)	A*D(1)	A*D(2)	A*B(1)
dep(1)	0.00113								
dep(2)	0.00091	0.00138							
chol(1)	0.00000	0.00001	0.00060						
chol(2)	0.00000	0.00002	0.00030	0.00060					
chol(3)	0.00000	0.00002	0.00030	0.00030	0.00060				
chol(4)	-0.00000	0.00002	0.00030	0.00030	0.00030	0.00059			
A*D(1)	-0.00113	-0.00091	0.00000	0.00000	-0.00000	0.00000	0.00220		
A*D(2)	-0.00090	-0.00138	-0.00000	0.00001	0.00000	0.00000	0.00179	0.00265	
A*B(1)	0.00004	0.00004	-0.00000	0.00000	-0.00000	0.00000	-0.00009	-0.00009	0.00183
A*B(2)	0.00007	0.00012	-0.00000	0.00001	0.00001	0.00002	-0.00014	-0.00024	0.00094
CONSTANT	-0.00086	-0.00086	-0.00030	-0.00029	-0.00029	-0.00028	0.00085	0.00084	0.00046

	A*B(2)	CONSTANT
A*B(2)	0.00149	
CONSTANT	0.00042	0.00160

6. CHD hospitalisation for the men aged 45-64 at screening visit 1

	chol(1)	chol(2)	chol(3)	chol(4)	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)
chol(1)	0.01140								
chol(2)	0.00621	0.00966							
chol(3)	0.00622	0.00624	0.00944						
chol(4)	0.00623	0.00625	0.00627	0.00852					
alc(1)	-0.00002	-0.00003	-0.00002	-0.00002	0.00145				
alc(2)	-0.00001	-0.00001	-0.00000	-0.00001	0.00099	0.00442			
alc(3)	0.00001	-0.00000	0.00003	0.00004	0.00099	0.00103	0.00843		
alc(4)	-0.00002	-0.00001	-0.00004	-0.00002	0.00099	0.00103	0.00104	0.01880	
alc(5)	-0.00002	0.00001	0.00001	-0.00006	0.00099	0.00103	0.00105	0.00104	0.02742
alc(6)	0.00014	0.00013	0.00018	0.00022	0.00099	0.00105	0.00107	0.00106	0.00108
age	0.00624	0.00616	0.00610	0.00595	0.00012	0.00018	0.00015	0.00007	0.00017
smok(1)	-0.00002	-0.00008	-0.00010	-0.00020	-0.00007	-0.00017	-0.00015	-0.00013	-0.00011
smok(2)	0.00000	-0.00009	-0.00013	-0.00027	-0.00005	-0.00017	-0.00023	-0.00021	-0.00024
bmi	-0.00007	-0.00012	-0.00015	-0.00019	0.00002	0.00001	0.00001	0.00002	-0.00000
dep(1)	0.00005	0.00007	0.00006	-0.00001	0.00004	0.00004	-0.00006	-0.00003	-0.00004
dep(2)	0.00019	0.00023	0.00023	0.00030	0.00010	0.00001	-0.00011	-0.00009	-0.00011
C*D(1)	0.00002	0.00009	0.00010	0.00021	-0.00004	-0.00001	-0.00002	-0.00003	-0.00005
C*D(2)	-0.00001	0.00007	0.00011	0.00024	-0.00006	-0.00005	0.00001	-0.00005	0.00004
dbp	-0.00009	-0.00018	-0.00024	-0.00031	-0.00005	-0.00013	-0.00022	-0.00015	-0.00018
C*G	0.00009	0.00019	0.00026	0.00033	0.00001	0.00004	0.00014	-0.00001	0.00010
C*A(1)	-0.01140	-0.00621	-0.00621	-0.00622	-0.00000	-0.00001	-0.00005	0.00003	-0.00001
C*A(2)	-0.00621	-0.00965	-0.00623	-0.00624	-0.00002	-0.00002	-0.00004	0.00001	-0.00007
C*A(3)	-0.00621	-0.00623	-0.00943	-0.00625	-0.00003	-0.00004	-0.00010	0.00005	-0.00002
C*A(4)	-0.00622	-0.00624	-0.00625	-0.00850	-0.00003	-0.00003	-0.00010	0.00000	0.00002
C*F(1)	-0.00005	-0.00007	-0.00005	0.00001	0.00001	0.00001	0.00003	0.00011	0.00001
C*F(2)	-0.00019	-0.00023	-0.00023	-0.00030	-0.00003	0.00001	0.00003	0.00009	-0.00003
CONSTANT	-0.00620	-0.00609	-0.00602	-0.00585	-0.00099	-0.00091	-0.00078	-0.00079	-0.00079

	alc(6)	age	smok(1)	smok(2)	bmi	dep(1)	dep(2)	C*D(1)	C*D(2)
alc(6)	0.02343								
age	0.00018	0.02495							
smok(1)	-0.00012	0.00417	0.00732						
smok(2)	-0.00030	0.00407	0.00451	0.00579					
bmi	0.00003	-0.00005	-0.00005	0.00012	0.00114				
dep(1)	-0.00008	0.00668	-0.00023	-0.00038	-0.00003	0.00815			
dep(2)	-0.00032	0.00674	-0.00023	-0.00065	-0.00002	0.00696	0.00933		
C*D(1)	-0.00007	-0.00612	-0.00731	-0.00450	0.00002	0.00023	0.00022	0.01067	
C*D(2)	0.00007	-0.00603	-0.00451	-0.00576	0.00004	0.00037	0.00065	0.00666	0.00889
dbp	-0.00023	0.00148	0.00001	0.00016	-0.00016	-0.00001	-0.00002	0.00001	-0.00017
C*G	0.00012	-0.00244	0.00000	-0.00017	0.00002	0.00001	0.00002	0.00002	0.00028

	alc(6)	age	smok(1)	smok(2)	bmi	dep(1)	dep(2)	C*D(1)	C*D(2)
C*A(1)	-0.00013	-0.00949	0.00002	-0.00000	0.00003	-0.00005	-0.00019	-0.00006	0.00000
C*A(2)	-0.00009	-0.00945	0.00009	0.00008	0.00005	-0.00007	-0.00023	-0.00012	-0.00005
C*A(3)	-0.00019	-0.00931	0.00010	0.00012	0.00006	-0.00005	-0.00024	-0.00017	-0.00009
C*A(4)	-0.00022	-0.00911	0.00021	0.00026	0.00008	0.00001	-0.00030	-0.00033	-0.00029
C*F(1)	0.00010	-0.01032	0.00023	0.00037	-0.00001	-0.00815	-0.00696	-0.00041	-0.00063
C*F(2)	0.00018	-0.01043	0.00023	0.00065	-0.00002	-0.00696	-0.00932	-0.00041	-0.00106
CONSTANT	-0.00084	-0.01646	-0.00410	-0.00408	-0.00049	-0.00669	-0.00679	0.00420	0.00411

	dbp	C*G	C*A(1)	C*A(2)	C*A(3)	C*A(4)	C*F(1)	C*F(2)	CONSTANT
dbp	0.00294								
C*G	-0.00291	0.00464							
C*A(1)	0.00009	-0.00014	0.01702						
C*A(2)	0.00019	-0.00026	0.00944	0.01498					
C*A(3)	0.00026	-0.00035	0.00944	0.00946	0.01460				
C*A(4)	0.00033	-0.00045	0.00945	0.00948	0.00949	0.01336			
C*F(1)	0.00001	0.00000	0.00010	0.00013	0.00004	-0.00003	0.01268		
C*F(2)	0.00003	-0.00001	0.00029	0.00040	0.00037	0.00047	0.01076	0.01452	
CONSTANT	-0.00137	0.00146	0.00623	0.00615	0.00609	0.00593	0.00668	0.00676	0.01727

7. Cancer hospitalisation in the men aged 45-64 at screening visit 1

	age	smok(1)	smok(2)	bmi	chol(1)	chol(2)	chol(3)	chol(4)	alc(1)
age	0.00213								
smok(1)	-0.00015	0.00370							
smok(2)	-0.00000	0.00225	0.00307						
bmi	0.00001	-0.00007	0.00024	0.00169					
chol(1)	-0.00002	-0.00004	-0.00002	-0.00007	0.00369				
chol(2)	-0.00002	-0.00005	-0.00002	-0.00013	0.00163	0.00376			
chol(3)	-0.00002	-0.00008	-0.00002	-0.00017	0.00163	0.00164	0.00389		
chol(4)	0.00000	-0.00014	-0.00010	-0.00023	0.00164	0.00165	0.00166	0.00378	
alc(1)	0.00009	-0.00014	-0.00014	0.00003	-0.00004	-0.00007	-0.00006	-0.00004	0.00238
alc(2)	0.00024	-0.00027	-0.00034	0.00001	-0.00003	-0.00003	-0.00004	-0.00003	0.00169
alc(3)	0.00031	-0.00026	-0.00037	0.00000	-0.00004	-0.00005	-0.00004	-0.00002	0.00169
alc(4)	0.00019	-0.00025	-0.00043	-0.00000	0.00001	0.00000	0.00001	0.00001	0.00169
alc(5)	0.00034	-0.00025	-0.00038	-0.00002	-0.00005	-0.00005	0.00001	-0.00003	0.00169
alc(6)	0.00040	-0.00026	-0.00042	0.00002	0.00008	0.00010	0.00008	0.00011	0.00168
dep(1)	-0.00007	-0.00018	-0.00028	-0.00005	0.00004	0.00005	0.00001	-0.00001	0.00006
dep(2)	-0.00012	-0.00019	-0.00043	-0.00004	0.00010	0.00015	0.00014	0.00016	0.00012
CONSTANT	-0.00156	-0.00176	-0.00192	-0.00073	-0.00159	-0.00158	-0.00153	-0.00148	-0.00167

	alc(2)	alc(3)	alc(4)	alc(5)	alc(6)	dep(1)	dep(2)	CONSTANT
alc(2)	0.00599							
alc(3)	0.00174	0.01143						
alc(4)	0.00174	0.00175	0.02352					
alc(5)	0.00175	0.00177	0.00176	0.02969				
alc(6)	0.00176	0.00179	0.00178	0.00181	0.02122			
dep(1)	0.00007	-0.00007	0.00007	-0.00006	-0.00003	0.00383		
dep(2)	0.00002	-0.00016	-0.00006	-0.00023	-0.00032	0.00311	0.00439	
CONSTANT	-0.00163	-0.00154	-0.00156	-0.00154	-0.00165	-0.00288	-0.00288	0.00833

8. Hospitalisation for injury or poisoning for men aged 45-64 at screening visit 1

	dep(1)	dep(2)	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)	alc(6)	smok(1)
dep(1)	0.01167								
dep(2)	0.01008	0.01236							
alc(1)	0.00008	0.00017	0.00350						
alc(2)	0.00011	0.00004	0.00259	0.00678					
alc(3)	-0.00003	-0.00015	0.00258	0.00263	0.01151				
alc(4)	0.00006	-0.00009	0.00259	0.00264	0.00266	0.01858			
alc(5)	-0.00005	-0.00017	0.00258	0.00263	0.00265	0.00266	0.02279		
alc(6)	-0.00003	-0.00034	0.00257	0.00265	0.00267	0.00269	0.00268	0.01435	
smok(1)	-0.00047	-0.00060	-0.00016	-0.00034	-0.00031	-0.00034	-0.00030	-0.00030	0.00954
smok(2)	-0.00073	-0.00120	-0.00017	-0.00049	-0.00056	-0.00066	-0.00061	-0.00065	0.00480
bmi	0.00946	0.00920	0.00000	-0.00012	-0.00016	-0.00022	-0.00023	-0.00017	0.00427
chol(1)	0.00004	0.00014	-0.00005	-0.00003	-0.00003	-0.00001	-0.00007	0.00011	-0.00006
chol(2)	0.00005	0.00019	-0.00007	-0.00004	-0.00005	-0.00001	-0.00005	0.00010	-0.00009
chol(3)	0.00005	0.00022	-0.00006	-0.00004	-0.00002	-0.00004	-0.00002	0.00011	-0.00014
chol(4)	0.00001	0.00023	-0.00005	-0.00004	-0.00000	-0.00002	-0.00008	0.00016	-0.00018
b*d(1)	-0.01167	-0.01006	0.00001	-0.00003	-0.00005	-0.00004	-0.00001	-0.00001	0.00048
b*d(2)	-0.01006	-0.01231	-0.00000	-0.00001	0.00002	0.00005	-0.00004	0.00002	0.00061
s*b(1)	0.00047	0.00060	0.00001	0.00008	0.00010	0.00014	0.00013	0.00009	-0.00952
s*b(2)	0.00073	0.00121	0.00004	0.00021	0.00026	0.00035	0.00034	0.00031	-0.00477
CONSTANT	-0.00955	-0.00945	-0.00253	-0.00228	-0.00211	-0.00213	-0.00204	-0.00212	-0.00403

	smok(2)	bmi	chol(1)	chol(2)	chol(3)	chol(4)	b*d(1)	b*d(2)	s*b(1)
smok(2)	0.00626								
bmi	0.00395	0.02695							
chol(1)	0.00001	-0.00008	0.00452						
chol(2)	0.00000	-0.00012	0.00200	0.00463					
chol(3)	0.00001	-0.00013	0.00201	0.00202	0.00475				
chol(4)	-0.00007	-0.00020	0.00201	0.00203	0.00205	0.00472			
b*d(1)	0.00074	-0.02129	0.00000	-0.00001	-0.00006	-0.00005	0.02579		
b*d(2)	0.00123	-0.02095	-0.00007	-0.00007	-0.00014	-0.00012	0.02224	0.02823	
s*b(1)	-0.00477	-0.00787	0.00004	0.00005	0.00009	0.00004	-0.00085	-0.00089	0.01757
s*b(2)	-0.00621	-0.00731	-0.00002	-0.00007	-0.00010	-0.00015	-0.00126	-0.00205	0.00869
CONSTANT	-0.00374	-0.01256	-0.00199	-0.00197	-0.00196	-0.00188	0.00948	0.00927	0.00420

	s*b(2)	CONST
s*b(2)	0.01293	
CONSTANT	0.00393	0.01602

9. Hospitalisation for attempted suicide in men aged 45-64 at screening visit 1

	age	smok(1)	smok(2)	bmi	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)
age	0.02255								
smok(1)	-0.00198	1.09111							
smok(2)	-0.00028	0.16742	0.33470						
bmi	0.00026	-0.00179	0.00131	0.02368					
alc(1)	0.00112	-0.00255	-0.00121	0.00034	0.03438				
alc(2)	0.00296	-0.00538	-0.00461	0.00009	0.02075	0.05172			
alc(3)	0.00362	-0.00414	-0.00455	0.00015	0.02076	0.02140	0.09896		
alc(4)	0.00274	-0.00426	-0.00541	0.00054	0.02075	0.02136	0.02156	0.22238	
alc(5)	0.00396	-0.00391	-0.00540	0.00008	0.02077	0.02149	0.02174	0.02169	0.35655
alc(6)	0.00461	-0.00414	-0.00479	0.00090	0.02074	0.02166	0.02198	0.02197	0.02219
dep(1)	-0.00064	0.16706	0.16690	-0.00151	0.00095	0.00047	-0.00067	-0.00037	-0.00088
dep(2)	-0.00146	0.16742	0.16718	-0.00291	0.00207	-0.00094	-0.00211	-0.00089	-0.00335
B*E(1)	-0.00039	-1.09038	-0.16704	0.00040	0.00064	0.00146	0.00080	0.00104	0.00015
B*E(2)	-0.00009	-0.16702	-0.33385	0.00145	-0.00044	0.00014	-0.00039	0.00057	0.00017

	age	smok(1)	smok(2)	bmi	alc(1)	alc(2)	alc(3)	alc(4)	alc(5)
B*E(3)	-0.00048	-1.09066	-0.16710	0.00207	0.00102	0.00304	0.00120	0.00119	0.00252
B*E(4)	0.00024	-0.16727	-0.33381	0.00356	-0.00069	0.00093	0.00022	-0.00056	0.00114
CONSTANT	-0.00941	-0.16340	-0.16591	-0.00929	-0.02109	-0.01853	-0.01746	-0.01741	-0.01697

	alc(6)	dep(1)	dep(2)	B*E(1)	B*E(2)	B*E(3)	B*E(4)	CONSTANT
alc(6)	0.10067							
dep(1)	-0.00060	0.26683						
dep(2)	-0.00373	0.16730	0.29319					
B*E(1)	0.00084	-0.26665	-0.16697	1.26148				
B*E(2)	-0.00047	-0.26675	-0.16712	0.26668	0.44832			
B*E(3)	0.00142	-0.16702	-0.29250	1.09036	0.16706	1.27170		
B*E(4)	-0.00084	-0.16717	-0.29273	0.16699	0.33416	0.29256	0.47747	
CONSTANT	-0.01709	-0.16660	-0.16606	0.16638	0.16661	0.16538	0.16571	0.18740

10. All-cause mortality in the 14,950 women screened

	age(1)	age(2)	age(3)	smok(1)	smok(2)	CONSTANT
age(1)	0.12143					
age(2)	0.10037	0.10790				
age(3)	0.10039	0.10062	0.10935			
smok(1)	-0.00045	-0.00099	-0.00152	0.02119		
smok(2)	-0.00009	0.00060	0.00255	0.00780	0.01724	
CONSTANT	-0.10024	-0.10048	-0.10131	-0.00681	-0.00925	0.10570

11. Cancer mortality in the 14,950 women screened

	age(1)	age(2)	age(3)	smok(1)	smok(2)	dep(1)	dep(2)	CONSTANT
age(1)	0.18518							
age(2)	0.14330	0.15912						
age(3)	0.14329	0.14375	0.16320					
smok(1)	-0.00106	-0.00219	-0.00336	0.04408				
smok(2)	-0.00031	0.00113	0.00469	0.01669	0.03662			
dep(1)	0.00037	-0.00106	-0.00087	0.00215	-0.00151	0.13507		
dep(2)	-0.00106	-0.00243	-0.00035	0.00095	-0.00325	0.12680	0.17541	
CONSTANT	-0.14307	-0.14220	-0.14404	-0.01643	-0.01757	-0.12603	-0.12456	0.27248

12. CHD mortality in the 14,950 women screened

	age(1)	age(2)	age(3)	smok(1)	smok(2)	a*s(1)	a*s(2)	a*s(3)	a*s(4)
age(1)	6.44741								
age(2)	5.49018	5.61559							
age(3)	5.49018	5.49018	5.53475						
smok(1)	5.49018	5.49018	5.49018	24.83957					
smok(2)	5.49018	5.49018	5.49018	5.49018	15.36703				
a*s(1)	-6.44741	-5.49018	-5.49018	-24.83957	-5.49018	26.79729			
a*s(2)	-5.49018	-5.61559	-5.49018	-24.83957	-5.49018	24.83957	25.04929		
a*s(3)	-5.49018	-5.49018	-5.53475	-24.83957	-5.49018	24.83957	24.83957	25.01127	
a*s(4)	-6.44741	-5.49018	-5.49018	-5.49018	-15.36703	6.44741	5.49018	5.49018	16.52495
a*s(5)	-5.49018	-5.61559	-5.49018	-5.49018	-15.36703	5.49018	5.61559	5.49018	15.36703

	age(1)	age(2)	age(3)	smok(1)	smok(2)	a*s(1)	a*s(2)	a*s(3)	a*s(4)
a*s(6)	-5.49018	-5.49018	-5.53475	-5.49018	-15.36703	5.49018	5.49018	5.53475	15.36703
CONSTANT	-5.49018	-5.49018	-5.49018	-5.49018	-5.49018	5.49018	5.49018	5.49018	5.49018

	a*s(5)	a*s(6)	CONSTANT
a*s(5)	15.54090		
a*s(6)	15.36703	15.61529	
CONSTANT	5.49018	5.49018	5.49018

13. All-cause mortality in the 13,559 men who reached screening visit 3

	age	smok(1)	smok(2)	visc(1)	visc(2)	fibr(1)	fibr(2)	fibr(3)	fibr(4)
age	0.01242								
smok(1)	-0.00080	0.15565							
smok(2)	0.00019	0.10142	0.13561						
visc(1)	-0.00005	-0.00041	0.00027	0.01840					
visc(2)	-0.00051	-0.00042	0.00010	0.01196	0.01951				
fibr(1)	-0.00069	0.10153	0.10139	-0.00220	-0.00198	0.24621			
fibr(2)	-0.00127	0.10161	0.10136	-0.00359	-0.00409	0.10203	0.18766		
fibr(3)	-0.00164	0.10168	0.10134	-0.00473	-0.00585	0.10224	0.10298	0.16824	
fibr(4)	-0.00157	0.10171	0.10134	-0.00513	-0.00882	0.10242	0.10349	0.10435	0.16455
dbp	-0.00004	0.00032	0.00027	-0.00042	-0.00097	-0.00005	-0.00018	-0.00020	-0.00056
s*f(1)	0.00023	-0.15561	-0.10144	0.00028	-0.00020	-0.24589	-0.10144	-0.10142	-0.10132
s*f(2)	0.00050	-0.10146	-0.13557	0.00012	0.00004	-0.24592	-0.10151	-0.10154	-0.10155
s*f(3)	0.00040	-0.15563	-0.10143	0.00085	0.00015	-0.10153	-0.18666	-0.10159	-0.10149
s*f(4)	0.00100	-0.10149	-0.13556	0.00010	0.00026	-0.10152	-0.18671	-0.10167	-0.10175
s*f(5)	0.00044	-0.15562	-0.10142	0.00071	-0.00022	-0.10151	-0.10151	-0.16617	-0.10136
s*f(6)	0.00097	-0.10147	-0.13556	-0.00005	-0.00038	-0.10148	-0.10150	-0.16617	-0.10147
s*f(7)	-0.00019	-0.15557	-0.10142	0.00070	0.00013	-0.10149	-0.10150	-0.10151	-0.16021
s*f(8)	0.00016	-0.10141	-0.13557	-0.00027	-0.00054	-0.10140	-0.10137	-0.10134	-0.16005
CONSTANT	-0.00779	-0.10091	-0.10181	-0.00741	-0.00593	-0.09996	-0.09874	-0.09786	-0.09711

	dbp	s*f(1)	s*f(2)	s*f(3)	s*f(4)	s*f(5)	s*f(6)	s*f(7)	s*f(8)
dbp	0.01010								
s*f(1)	-0.00047	0.36841							
s*f(2)	0.00026	0.24587	0.31194						
s*f(3)	-0.00051	0.15564	0.10144	0.31377					
s*f(4)	0.00054	0.10142	0.13565	0.18653	0.24128				
s*f(5)	-0.00012	0.15563	0.10145	0.15565	0.10145	0.26768			
s*f(6)	0.00074	0.10142	0.13565	0.10144	0.13570	0.16614	0.22147		
s*f(7)	0.00008	0.15559	0.10143	0.15561	0.10142	0.15562	0.10143	0.24261	
s*f(8)	0.00086	0.10140	0.13562	0.10140	0.13564	0.10144	0.13567	0.16019	0.20521
CONSTANT	-0.00564	0.10150	0.10092	0.10118	0.10041	0.10103	0.10045	0.10127	0.10099

	CONSTANT
CONSTANT	0.11298

14. Cancer mortality in the 13,559 men who reached screening visit 3

	age	smok(1)	smok(2)	visc(1)	visc(2)	CONSTANT
age	0.05386					
smok(1)	-0.00215	0.09328				
smok(2)	0.00139	0.05951	0.07224			
visc(1)	-0.00110	-0.00017	-0.00147	0.05662		
visc(2)	-0.00367	-0.00261	-0.00470	0.03505	0.05473	
CONSTANT	-0.04259	-0.05660	-0.05831	-0.03310	-0.02857	0.11668

15. CHD mortality in the 13,559 men who reached screening visit 3

	age	smok(1)	smok(2)	dbp	visc(1)	visc(2)	hdl(1)	hdl(2)	CONSTANT
age	0.02505								
smok(1)	-0.00153	0.05950							
smok(2)	0.00089	0.03907	0.04834						
dbp	-0.00011	0.00015	0.00167	0.02423					
visc(1)	-0.00076	-0.00002	-0.00104	-0.00091	0.03865				
visc(2)	-0.00259	-0.00172	-0.00302	-0.00208	0.02061	0.03284			
hdl(1)	-0.00054	0.00030	0.00201	0.00014	0.00044	0.00172	0.02823		
hdl(2)	-0.00119	0.00041	0.00289	-0.00016	0.00095	0.00174	0.01157	0.03654	
CONSTANT	-0.01547	-0.03750	-0.04069	-0.01607	-0.01906	-0.01622	-0.01330	-0.01341	0.08093

G: Tabulations to investigate interaction effects for men age 45-64

These tabulations relate to the significant interactions in the models fitted for the 73,110 men aged 45-64 at screening visit 1.

1. For all-cause mortality the tabulations of the observed and expected proportions of deaths/subjects for the DEPCAT*smoking interaction were as follows.

Observed	Never smoked	Ex-smoker	Current smoker
Affluent	0.0220	0.0231	0.0433
Middle	0.0237	0.0442	0.0569
Deprived	0.0324	0.0617	0.0744

Expected	Never smoked	Ex-smoker	Current smoker
Affluent	0.0188	0.0321	0.0457
Middle	0.0267	0.0448	0.0626
Deprived	0.0339	0.0571	0.0797

2. For all-cause mortality the tabulations of the observed and expected proportions of deaths/subjects for the BMI*smoking interaction were as follows.

Observed	Never smoked	Ex-smoker	Current smoker
BMI < 25.46 kg/m^2	0.0231	0.0514	0.0670
BMI ≥ 25.46 kg/m^2	0.0272	0.0411	0.0535

Expected	Never smoked	Ex-smoker	Current smoker
BMI < 25.46 kg/m^2	0.0290	0.0489	0.0688
BMI ≥ 25.46 kg/m^2	0.0249	0.0420	0.0580

3. For all-cause mortality the tabulations of the observed and expected proportions of deaths/subjects for the age*smoking interaction were as follows.

Observed	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.0116	0.0189	0.0341
Age 55-64	0.0407	0.0655	0.0916

Expected	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.0140	0.0233	0.0333
Age 55-64	0.0410	0.0676	0.0948

4. For all-cause mortality the tabulations of the observed and expected proportions of deaths/subjects for the age*DBP interaction were as follows.

Observed	DBP < 84 mmHg	DBP ≥ 84 mmHg
Age 45-54	0.0220	0.0256
Age 55-64	0.0696	0.0701

Expected	DBP < 84 mmHg	DBP ≥ 84 mmHg
Age 45-54	0.0225	0.0249
Age 55-64	0.0654	0.0723

5. For CHD mortality the tabulations of the observed and expected proportions of deaths/subjects for the age*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.0043	0.0073	0.0134
Age 55-64	0.0167	0.0264	0.0327

Expected	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.0052	0.0084	0.0115
Age 55-64	0.0145	0.0234	0.0320

Contrasts	Coeff / s.e.
Age45-54(Ex-Never smoked) - Age55-64(Ex-Never smoked)	-0.788
Age45-54(Current-Never smoked) - Age55-64(Current-Never smoked)	-2.80

6. For CHD mortality the tabulations of the observed and expected proportions of deaths/subjects for the DBP*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
DBP < 84 mmHg	0.0093	0.0177	0.0185
DBP ≥ 84 mmHg	0.0109	0.0185	0.0268

Expected	Never smoked	Ex-smoker	Current smoker
DBP < 84 mmHg	0.0085	0.0143	0.0194
DBP ≥ 84 mmHg	0.0106	0.0174	0.0237

7. For CHD mortality the tabulations of the observed and expected proportions of deaths/subjects for the alcohol*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
0 units alcohol / week	0.0139	0.0302	0.0281
1-20 units / week	0.0088	0.0161	0.0223
21-30 units / week	0.0087	0.0138	0.0149
31-40 units / week	0.0070	0.0038	0.0228
41-50 units / week	0.0036	0.0077	0.0243
51-60 units / week	0.0114	0.0171	0.0242
>60 units / week	0.0050	0.0140	0.0222

Expected	Never smoked	Ex-smoker	Current smoker
0 units alcohol / week	0.0130	0.0212	0.0289
1-20 units / week	0.0093	0.0152	0.0208
21-30 units / week	0.0074	0.0121	0.0166
31-40 units / week	0.0082	0.0136	0.0184
41-50 units / week	0.0086	0.0146	0.0196
51-60 units / week	0.0111	0.0190	0.0256
>60 units / week	0.0097	0.0162	0.0218

8. For CHD mortality the tabulations of the observed and expected proportions of deaths/subjects for the DEPCAT*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
Affluent	0.0085	0.0080	0.0160
Middle	0.0098	0.0174	0.0218
Deprived	0.0127	0.0263	0.0259

Expected	Never smoked	Ex-smoker	Current smoker
Affluent	0.0066	0.0111	0.0150
Middle	0.0095	0.0155	0.0213
Deprived	0.0123	0.0203	0.0266

9. For all-cause hospitalisation the tabulations of the observed and expected proportions of deaths/subjects for the age*DEPCAT interaction were as follows:

Observed	Affluent	Middle	Deprived
Age 45-54	0.2635	0.3010	0.3551
Age 55-64	0.3678	0.4211	0.4455

Expected	Affluent	Middle	Deprived
Age 45-54	0.2718	0.3064	0.3385
Age 55-64	0.3711	0.4123	0.4487

10. For all-cause hospitalisation the tabulations of the observed and expected proportions of deaths/subjects for the age*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.2624	0.3045	0.3414
Age 55-64	0.3721	0.4314	0.4434

Expected	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.2680	0.3167	0.3345
Age 55-64	0.3673	0.4233	0.4436

11. For CHD hospitalisation the tabulations of the observed and expected proportions of deaths/subjects for the age*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.0213	0.0423	0.0522
Age 55-64	0.0520	0.0774	0.0748

Expected	Never smoked	Ex-smoker	Current smoker
Age 45-54	0.0228	0.0374	0.0412
Age 55-64	0.0384	0.0612	0.0676

12. For CHD hospitalisation the tabulations of the observed and expected proportions of deaths/subjects for the age*DBP interaction were as follows:

Observed	DBP < 84 mmHg	DBP ≥ 84 mmHg
Age 45-54	0.0357	0.0458
Age 55-64	0.0677	0.0713

Expected	DBP < 84 mmHg	DBP ≥ 84 mmHg
Age 45-54	0.0324	0.0355
Age 55-64	0.0539	0.0588

13. For CHD hospitalisation the tabulations of the observed and expected proportions of deaths/subjects for the age*cholesterol interaction were as follows:

Observed	Age 45-54	Age 55-64
Chol < 4.94 mmol/l	0.0223	0.0445
4.94-5.55 mmol/l	0.0275	0.0612
5.56-6.11 mmol/l	0.0423	0.0703
6.12-6.80 mmol/l	0.0461	0.0768
>6.80 mmol/l	0.0655	0.0957

Expected	Age 45-54	Age 55-64
Chol < 4.94 mmol/l	0.0206	0.0352
4.94-5.55 mmol/l	0.0273	0.0456
5.56-6.11 mmol/l	0.0351	0.0578
6.12-6.80 mmol/l	0.0379	0.0623
>6.80 mmol/l	0.0484	0.0805

14. For CHD hospitalisation the tabulations of the observed and expected proportions of deaths/subjects for the age*DEPCAT interaction were as follows:

Observed	Affluent	Middle	Deprived
Age 45-54	0.0258	0.0414	0.0484
Age 55-64	0.0571	0.0698	0.0753

Expected	Affluent	Middle	Deprived
Age 45-54	0.0279	0.0339	0.0392
Age 55-64	0.0470	0.0558	0.0647

15. For hospitalisation for injury or poisoning the tabulations of the observed and expected proportions of deaths/subjects for the BMI*DEPCAT interaction were as follows:

Observed	Affluent	Middle	Deprived
BMI < 25.46 kg/m^2	0.0185	0.0300	0.0484
BMI ≥ 25.46 kg/m^2	0.0161	0.0242	0.0281

Expected	Affluent	Middle	Deprived
BMI < 25.46 kg/m^2	0.0299	0.0415	0.0449
BMI ≥ 25.46 kg/m^2	0.0298	0.0349	0.033

16. For hospitalisation for injury or poisoning the tabulations of the observed and expected proportions of deaths/subjects for the BMI*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
BMI < 25.46 kg/m^2	0.0236	0.0258	0.0416
BMI ≥ 25.46 kg/m^2	0.0238	0.0204	0.0277

Expected	Never smoked	Ex-smoker	Current smoker
BMI < 25.46 kg/m^2	0.0390	0.0388	0.0386
BMI ≥ 25.46 kg/m^2	0.0348	0.0370	0.0382

17. For hospitalisation for attempted suicide the tabulations of the observed and expected proportions of deaths/subjects for the DEPCAT*smoking interaction were as follows:

Observed	Never smoked	Ex-smoker	Current smoker
Affluent	0.0015	0.0009	0.0018
Middle	0.0003	0.0011	0.0036
Deprived	0.0017	0.0038	0.0053

Expected	Never smoked	Ex-smoker	Current smoker
Affluent	0.0010	0.0013	0.0028
Middle	0.0017	0.0023	0.0047
Deprived	0.0026	0.0036	0.0074

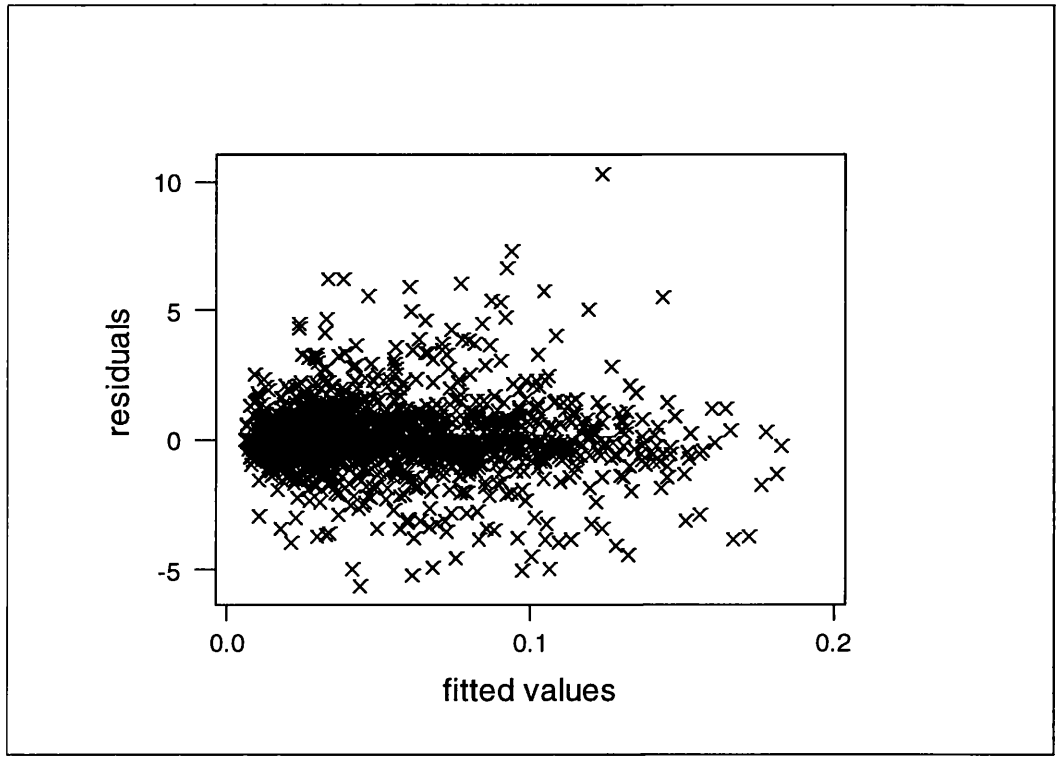
H: Logistic model checking

While model checking was carried out for each model fitted, only one example is given here. The results for each model were similar.

Plot of residuals against fitted values

These plots of residuals against fitted values were constructed to investigate model fit. Large residuals indicate observations which are poorly accounted for by the model. Raw residuals (the difference between the observed and predicted values) were used here, although the chi residuals (the square root of the contribution for the observation to the Pearson chi-square) and deviance residuals (the square root of the deviance contribution for the observation, with sign equal to the sign of the residual) could also have been used (SAS Institute Inc., 1993).

This residual plot relates to the final model for all-cause mortality in middle-aged men at screening visit 1, in which interaction terms were included. It shows that while there may be one outlying observation, most points are reasonably accounted for by the fitted model.



Dispersion parameters

The Pearsons chi-square statistic, $\chi^2 = \frac{\sum_i w_i (y_i - \mu_i)^2}{V(\mu_i)}$, and its scaled version, $\frac{\chi^2}{\phi}$ (distributed as $\chi^2(n-p)$ under the null hypothesis), where ϕ is the dispersion parameter, can be used as an approximate guide to the goodness of fit of a given model (SAS Institute Inc., 1993). The chi-square statistic and it's degrees of freedom can be used to give an estimate of the dispersion parameter $\hat{\phi} = \frac{\chi^2}{n-p}$. For a good fitting model, the dispersion parameter should be close to one. The data may be overdispersed if the dispersion estimate is greater than one or underdispersed if the dispersion estimate is less than one.

For each of the models fitted, the dispersion parameter was close to one, indicating no problems with model fit. In the example illustrated above for all-cause mortality in middle-aged men at screening visit 1, χ^2=2201.5 with 2245 degrees of freedom, leading to $\hat{\phi}$=0.981, which is clearly close to one.

I: Postcode districts making up the screening area

The following postcode districts have been used to define the area in which screening was carried out for WOSCOPS. They are divided into five postcode areas.

1. North Glasgow

G1, G2, G3, G4, G11, G12, G13, G14, G15, G20, G21, G22, G23, G31, G32, G33, G34, G60, G61, G62, G63, G64, G65, G66, G67, G68, G69, G81, G82, G83, G84.

2. South Glasgow

G5, G40, G41, G42, G43, G44, G45, G46, G51, G52, G53, G71, G72, G73, G76, G77, G78.

3. East Kilbride and Lanarkshire

G74, G75, ML1, ML2, ML3, ML4, ML5, ML6, ML7, ML8, ML9, ML10, ML11, ML12.

4. Renfrew, Paisley and Greenock

PA1, PA2, PA3, PA4, PA5, PA6, PA7, PA8, PA9, PA10, PA11, PA12, PA13, PA14, PA15, PA16, PA17, PA18, PA19.

5.Dumfries and Galloway

DG1, DG2, DG5, DG6, DG7, DG8, DG9, DG10, DG11, DG12.

J: Parameter estimates from stepwise models

Parameter estimates are given here for the contrasts in the final models. The coefficient divided by the standard error (SE) of the coefficient gives an approximate z-statistic, and the exponential of the coefficient gives the estimated odds ratio for each contrast.

1. The final model for all-cause mortality in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 /45-54)	1.38	0.12	12.0	3.99
smok(Ex/Never)	0.13	0.20	0.65	1.14
(current/never)	1.08	0.17	6.43	2.93
depcat(middle/affluent)	-0.00	0.12	-0.01	1.0
(deprived/affluent)	0.30	0.14	2.19	1.35
alcohol(1-20 / 0)	-0.31	0.04	-7.30	0.74
(21-30 / 0)	-0.42	0.07	-6.02	0.66
(31-40 / 0)	-0.28	0.10	-2.88	0.76
(41-50 / 0)	-0.20	0.12	-1.66	0.82
(51-60 / 0)	0.06	0.14	0.44	1.06
(>60 / 0)	0.05	0.12	0.42	1.05
cholesterol(q2/q1)	-0.29	0.06	-5.16	0.75
(q3/q1)	-0.22	0.06	-3.95	0.80
(q4/q1)	-0.27	0.06	-4.83	0.76
(q5/q1)	-0.04	0.05	-0.75	0.96
bmi(high/low)	0.09	0.09	1.0	1.10
dbp(high/low)	0.25	0.07	3.60	1.29
middle(ex/never) / affluent(ex/never smoked)	0.64	0.18	3.58	1.90
middle(current/never) / affluent(current/never smoked)	0.24	0.15	1.56	1.27
deprived(ex/never) / affluent(ex/never smoked)	0.65	0.19	3.39	1.92
deprived(current/never) / affluent(current/never smoked)	0.17	0.16	1.06	1.19
exsmoker(high/low BMI) / never smoked(high/low BMI)	-0.34	0.11	-2.95	0.71
current smoker(high/low BMI) / never smokedhigh/low BMI)	-0.33	0.10	-3.15	0.72
age55-64(high/low DBP) / age45-54(high/low DBP)	-0.18	0.08	-2.25	0.83
age55-64(ex/never) / age45-54(ex/never smoked)	-0.00	0.14	-0.01	1.00
age45-54(current/never) / age55-65(current/never smoked)	-0.27	0.12	-2.33	0.76
CONSTANT	-4.36	0.15	-28.8	0.01

2. The final model for CHD mortality in males age 45-64.

Main effects contrasts:

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	1.42	0.18	8.10	4.13
smok(Ex/Never)	0.16	0.36	0.44	1.17
(current/never)	0.78	0.30	2.58	2.18
depcat(middle/affluent)	0.04	0.20	0.20	1.04
(deprived/affluent)	0.33	0.22	1.51	1.39
alcohol(1-20 / 0)	-0.36	0.15	-2.35	0.70
(21-30 / 0)	-0.29	0.30	-0.96	0.75
(31-40 / 0)	-0.46	0.52	-0.90	0.63
(41-50 / 0)	-1.25	1.01	-1.24	0.29
(51-60 / 0)	-0.03	0.72	-0.04	0.97
(>60 / 0)	-0.81	1.01	-0.80	0.44
cholesterol(q2/q1)	0.03	0.10	0.34	1.03
(q3/q1)	0.20	0.10	2.09	1.22
(q4/q1)	0.23	0.09	2.42	1.26
(q5/q1)	0.60	0.09	6.84	1.82
dbp(high/low)	0.06	0.14	0.45	1.07

Interaction contrasts:

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
affluent (ex/never) / middle(ex/never smoked)	0.69	0.29	2.36	1.99
affluent(current/never) / middle(current/never smoked)	0.24	0.24	0.99	1.27
affluent(ex/never) / deprived(ex/never smoked)	0.83	0.31	2.65	2.29
affluent(current/never) / deprived(current/never smoked)	0.15	0.26	0.56	1.16
never smoked(high/low DBP) / exsmoker(high/low DBP)	-0.05	0.18	-0.30	0.95
never smoked(high/low DBP) / current smoker(high/low DBP)	0.28	0.16	1.72	1.32
exsmoker(1-20 / 0) / never smoker(1-20 / 0)	-0.19	0.19	-0.99	0.83
exsmoker(21-30 / 0) / never smoker(21-30 / 0)	0.16	0.18	0.88	1.17
exsmoker(31-40 / 0) / never smoker(31-40 / 0)	-0.33	0.37	-0.89	0.72
exsmoker(41-50 / 0) / never smoker(41-50 / 0)	-0.28	0.34	-0.81	0.76
exsmoker(51-60 / 0) / never smoker(51-60 / 0)	-1.54	0.78	-1.97	0.21
exsmoker(>60 / 0) / never smoker(>60 / 0)	0.36	0.55	0.66	1.44
current smoker(1-20 / 0) / never smoker(1-20 / 0)	-0.03	1.17	-0.03	0.97
current smoker(21-30 /0) / never smoker(21-30 / 0)	1.16	1.04	1.12	3.18
current smoker(31-40 / 0) / never smoker(31-40 / 0)	-0.47	0.89	-0.53	0.62
current smoker(41-50 / 0) / never smoker(41-50 / 0)	0.06	0.78	0.07	1.01
current smoker(51-60 / 0) / never smoker(51-60 / 0)	0.08	1.14	0.07	1.09
current smoker(>60 / 0) / never smoker(>60 / 0)	0.71	1.04	0.68	2.03
age45-54(ex/never) / age55-64(ex/never smoked)	-0.18	0.22	-0.79	0.84
age45-54(current/never) / age55-65(current/never smoked)	-0.54	0.19	-2.80	0.58
CONSTANT	-5.61	0.26	-21.5	0.00

3. Cancer mortality in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	1.19	0.07	17.4	3.30
smoking(ex/never smoked)	0.54	0.10	5.51	1.71
(current/never smoked)	0.96	0.09	10.9	2.61
bmi(high/low)	-0.39	0.06	-6.31	0.68
cholesterol(q2/q1)	-0.38	0.09	-4.30	0.69
(q3/q1)	-0.31	0.09	-3.56	0.74
(q4/q1)	-0.48	0.09	-5.31	0.62
(q5/q1)	-0.43	0.09	-4.78	0.65
depcat(middle/affluent)	0.20	0.10	1.96	1.22
(deprived/affluent)	0.37	0.10	3.57	1.45
CONSTANT	-5.26	0.13	-39.3	0.01

4. Trauma mortality in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	0.49	0.18	2.66	1.64
alcohol(1-20 / 0)	-0.22	0.21	-1.01	0.80
(21-30 / 0)	-0.58	0.40	-1.46	0.56
(31-40 / 0)	0.05	0.45	0.12	1.06
(41-50 / 0)	-0.41	0.73	-0.57	0.66
(51-60 / 0)	1.23	0.45	2.74	3.41
(>60 / 0)	0.26	0.61	0.43	1.30
CONSTANT	-6.50	0.22	-29.9	0.00

5. All-cause hospitalisation in males age 45-54.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	0.48	0.05	10.0	1.62
smoking(ex/never)	0.21	0.03	6.61	1.24
(current/never)	0.35	0.03	12.6	1.42
alcohol(1-20 / 0)	-0.24	0.02	-12.4	0.79
(21-30 / 0)	-0.28	0.03	-9.40	0.76
(31-40 / 0)	-0.24	0.04	-5.66	0.79
(41-50 / 0)	-0.34	0.06	-6.01	0.71
(51-60 / 0)	-0.11	0.07	-1.54	0.90
(>60 / 0)	-0.05	0.06	-0.79	0.95
depcat(middle/affluent)	0.14	0.03	4.30	1.16
(deprived/affluent)	0.36	0.04	9.61	1.43
cholesterol(q2/q1)	-0.10	0.02	-3.98	0.91
(q3/q1)	-0.08	0.20	-3.23	0.92
(q4/q1)	-0.07	0.02	-2.83	0.93
(q5/q1)	0.02	0.02	1.01	1.03
age55-64(middle/affluent) / age45-54(middle/affluent)	0.04	0.05	0.92	1.04
age55-64(deprived/affluent) / age45-54(deprived/affluent)	-0.09	0.05	-1.66	0.92
age55-64(ex/never smoked)/ age45-54(ex/never smoker)	0.04	0.04	1.01	1.04
age55-64(current/never smoked) / age45-54(current/never smoked)	-0.06	0.04	-1.48	0.94
CONSTANT	-0.96	0.04	-24.1	0.38

6. CHD hospitalisation in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
cholesterol(q2/q1)	0.21	0.11	1.97	1.23
(q3/q1)	0.62	0.10	6.36	1.87
(q4/q1)	0.69	0.10	7.15	2.00
(q5/q1)	1.03	0.09	11.1	2.79
alcohol(1-20 / 0)	-0.33	0.04	-8.63	0.72
(21-30 / 0)	-0.62	0.07	-9.30	0.54
(31-40 / 0)	-0.49	0.09	-5.31	0.61
(41-50 / 0)	-0.75	0.14	-5.50	0.47
(51-60 / 0)	-0.62	0.17	-3.75	0.54
(>60 / 0)	-0.71	0.15	-4.67	0.49
age(55-64 / 45-54)	1.27	0.16	8.04	3.56
smoking(ex/never)	0.67	0.08	7.81	1.95
(current/never)	0.92	0.08	12.1	2.52
bmi(high/low)	0.27	0.03	7.87	1.30
depcat(middle/affluent)	0.40	0.09	4.39	1.49
(deprived/affluent)	0.56	0.10	5.83	1.76
age55-64(ex/never smoked) / age45-54(ex/never smoked)	-0.25	0.10	-2.38	0.78
age55-64(current/never smoked)/ age45-54(current/never smoked)	-0.47	0.09	-4.96	0.63
dbp(high/low)	0.21	0.05	3.93	1.24
age55-64(high/low DBP)/ age45-54(high/low DBP)	-0.18	0.07	-2.66	0.83
age55-64(q2/q1)/ age45-54(q2/q1)	0.12	0.13	0.94	1.13
age55-64(q3/q1)/ age45-54(q3/q1)	-0.14	0.12	-1.13	0.87
age55-64(q4/q1)/ age45-54(q4/q1)	-0.12	0.12	-1.01	0.89
age55-64(q5/q1)/ age45-54(q5/q1)	-0.23	0.12	-1.97	0.80
age55-64(middle/affluent)/ age45-54(middle/affluent))	-0.24	0.1	-2.14	0.78
age55-64(deprived/affluent)/ age45-54(deprived/affluent)	-0.29	0.12	-2.40	0.75
CONSTANT	-4.72	0.13	-35.9	0.00

7. Cancer hospitalisation in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	1.14	0.05	24.7	3.14
smoking(ex/never)	0.38	0.06	6.22	1.46
(current/never)	0.65	0.06	11.7	1.91
	-0.19	0.04	-4.65	0.83
bmi(high/low)				
cholesterol(q2/q1)	-0.24	0.06	-3.97	0.79
(q3/q1)	-0.26	0.06	-4.25	0.77
(q4/q1)	-0.32	0.06	-5.04	0.73
(q5/q1)	-0.25	0.06	-4.01	0.78
alcohol(1-20 / 0)	-0.13	0.05	-2.58	0.88
(21-30 / 0)	-0.15	0.08	-1.96	0.86
(31-40 / 0)	-0.04	0.11	-0.35	0.96
(41-50 / 0)	-0.28	0.15	-1.86	0.75
(51-60 / 0)	0.08	0.17	0.49	1.09
(>60 / 0)	0.15	0.15	1.03	1.16
depcat(middle/affluent)	-0.04	0.06	-0.73	0.96
(deprived/affluent)	0.12	0.07	1.82	1.13
CONSTANT	-4.06	0.09	-44.5	0.02

8. Trauma hospitalisation in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
depcat(middle/affluent)	0.43	0.11	3.91	1.53
(deprived/affluent)	0.82	0.11	7.33	2.26
alcohol(1-20 / 0)	0.26	0.06	0.44	1.03
(21-30 / 0)	0.23	0.08	2.79	1.26
(31-40 / 0)	0.35	0.11	3.27	1.42
(41-50 / 0)	0.42	0.14	3.08	1.52
(51-60 / 0)	0.72	0.15	4.76	2.05
(>60 / 0)	0.97	0.12	8.07	2.63
smoking(ex/never)	0.02	0.10	0.24	1.02
(current/never)	0.40	0.08	5.14	1.50
bmi(high/low)	0.06	0.16	0.38	1.06
cholesterol(q2/q1)	-0.20	0.07	-2.95	0.82
(q3/q1)	-0.22	0.07	-3.23	0.80
(q4/q1)	-0.24	0.07	-3.51	0.78
(q5/q1)	-0.22	0.07	-3.14	0.81
high BMI(middle/affluent)/ low BMI(middle/affluent)	-0.02	0.16	-0.14	0.98
high BMI(deprived/affluent)/ low BMI(deprived/affluent)	-0.32	0.17	-1.88	0.73
exsmoker(high/low BMI)/ never smoker(high/low BMI)	-0.20	0.13	-1.50	0.82
current smoker(high/low BMI)/ never smoker(high/low BMI)	-0.30	0.11	-2.66	0.74
CONSTANT	-4.06	0.13	-32.1	0.02

9. Hosptalisation for attempted suicide in males age 45-64.

TERM	COEFFICIENT	SE	COEF/SE	ODDS RATIO
age(55-64 / 45-54)	-0.33	0.15	-2.19	0.72
smoking(ex/never)	-1.47	1.04	-1.40	0.23
(current/never)	0.04	0.58	0.07	1.04
bmi(high/low)	-0.44	0.15	-2.86	0.64
alcohol(1-20 / 0)	-0.64	0.18	-3.48	0.52
(21-30 / 0)	0.18	0.23	0.78	1.19
(31-40 / 0)	0.07	0.32	0.23	1.08
(41-50 / 0)	-0.23	0.47	-0.49	0.80
(51-60 / 0)	-0.26	0.60	-0.43	0.78
(>60 / 0)	0.83	0.32	2.62	2.30
depcat(middle/affluent)	-0.53	0.52	-1.03	0.59
(deprived/affluent)	0.15	0.54	0.28	1.16
exsmoker(middle/affluent)/ never smoker(middle/affluent)	1.79	1.12	1.59	5.97
exsmoker(deprived/affluent)/ never smoker(deprived/affluent)	1.36	0.67	2.02	3.88
current smoker(middle/affluent)/ never smoker(middle/affluent)	2.21	1.13	1.96	9.13
current smoker(deprived/affluent)/ never smoker(deprived/affluent)	0.93	0.69	1.34	2.53
CONSTANT	-5.85	0.43	-13.5	0.00

References

Acheson ED (1987), Introduction. In Textbook of Medical Record Linkage (eds Baldwin JA, Acheson ED and Graham WJ), Oxford: Oxford University Press, pp 1-11.

Alderson MR, Bayliss RIS, Clarke CA, Whitfield AGW (1983), Death certification. British Medical Journal, 287, 444-445.

Arellano MG, Petersen GR, Petitti DB and Smith RE (1984), The California automated mortality linkage system (CAMLIS). American Journal of Public Health, 74, 1324-1330.

Arellano MG (1992), Comment. Journal of the American Statistical Society, 87, 1204-1206.

Black D, Morris JN, Smith C, Townsend P and Whitehead M (1988), Inequalities in health: The Black report; The health divide. Penguin Group, London.

Blackwelder WC, Yano K, Rhoads CG, Kagan A, Gordon T and Palesch Y (1980), Alcohol and mortality: The Honolulu Heart Study. American Journal of Medicine, 68, 164-169.

BMDP Statistical Software Inc. (1990), BMDP statistical software manual: to accompany the 1990 software release. Volume 2 pp 1021.

Buchwald H (1992), Cholesterol inhibition, cancer and chemotherapy, Lancet, 339, 1154-1156.

Budd D and Ginsberg H (1986), Hypocholesterolaemia and acute myelogenous leukaemia. Association between disease activity and plasma low-density lipoprotein cholesterol concentrations. Cancer, 58, 1361-1365.

Bush TL, Barrett-Connor E, Cowan LD, Criqui MH, Wallace RB, Suchindran CM et al (1987), Cardiovascular mortality and noncontraceptive use of oestrogen in women: results from the Lipid Research Clinics Program Follow-up Study. Circulation, 75, 1102-1109.

Carstairs V and Morris R (1991), Deprivation and health in Scotland. Aberdeen: Aberdeen University Press.

Chen Z, Peto R, Collins R, MacMahon S, Lu J and Li W (1991), Serum cholesterol concentration and coronary heart disease in population with low cholesterol concentrations. British Medical Journal, 303, 276-282.

Chyou PH, Nomura AMY, Stemmermann GN and Kato I (1992), Prospective study of serum cholesterol and site-specific cancers. Journal of Clinical Epidemiology, 45, 287-292.

Committee of Principal Investigators (1978), A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. British Heart Journal, 40, 1069-1118.

Committee of Principal Investigators (1984), WHO co-operative trial on primary prevention of ischaemic heart disease with clofibrate to lower serum cholesterol: Final mortality follow-up. (Report of the committee of principal investigators). Lancet, ii, 600-604.

Common Services Agency (1990), Scottish Cancer Registration Scheme (SMR6): scheme manual. Edinburgh: Common Services Agency.

Common Services Agency (1990), Scottish Cardiac Surgery Register: instruction manual. Edinburgh: Common Services Agency.

Copas JB, Hilton FJ (1990), Record linkage: Statistical models for matching computer records. Journal of the Royal Statistical Society-A, 153,3, 287-320.

Cowan LD, O'Connell DL, Criqui MH, Barrett-Connor E, Bush TL, Wallace RB (1990), Cancer mortality and lipid and lipoprotein levels. The Lipid Research Clinics Program Mortality Follow-up Study. American Journal of Epidemiology, 131, 468-482.

Crombie IK, Kenicer MB, Smith WCS, Tunstall-Pedoe HD, (1989), Unemployment, socioenvironmental factors, and coronary heart disease in Scotland. British Heart Journal, 61, 172-177.

Darne B, Girerd X, Safar M, Cambien F and Guize L (1989), Pulsatile versus steady component of blood pressure: a cross-sectional analysis and a prospective analysis on cardiovascular mortality. Hypertension, 13, 392-400.

Denholm SW, Macintyre CCA, Wilson JA, (1993), Audit of the Scottish Morbidity Record 1 (SMR1) Returns from an ENT unit. Health Bulletin, 51(6), 366-369.

Doll R, Peto R, Hall E, Wheatley K and Gray R (1994), Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors. British Medical Journal, 309, 911-918.

Donahue RP, Abbott RD, Reed DM and Yano KC (1988), Physical activity and coronary heart disease in middle-aged and elderly men. American Journal of Public Health, 78, 683-685.

Elford J, Phillips AN, Thomson AG and Shaper AG (1989), Migration and geographic variations in ischaemic heart disease in Great Britain. Lancet, 1, 343-346.

Elwood PC, Burr ML and Sweetnam PM (1992), Fish, fibre and heart disease. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 140-151.

Engelberg H, (1992), Low serum cholesterol and suicide. Lancet, 339, 727-729.

Farr W (1861), Report on army medical statistics. Parliamentary Paper 366.

Feinleib M (1981), On a possible inverse relationship between serum cholesterol and cancer mortality. American Journal of Epidemiology, 114, 5-10.

Fellegi IP, Sunter AB (1969), A theory for record linkage. Journal of the American Statistical Association, 64, 1183-1210.

Frank JW, Reed DM, Grove JS and Benfante R, (1992), Will lowering population levels of serum cholesterol affect total mortality? Expectations from the Honolulu Heart Program. Journal of Clinical Epidemiology, 45, 333-346.

Frick MH, Elo E, Haapa K, Heinonen OP, Heinsalmi P, Helo P, Huttunen JK, Kaitaniemi P, Koskinen P, Manninen V et al, (1987), Helsinki Heart Study: Primary-Prevention Trial with Gemfibrozil in Middle-Aged Men with Dyslipidemia: Safety of Treatment, Changes in Risk Factors, and Incidence of Coronary Heart Disease. New England Journal of Medicine, 317,1237-1245.

Friedman LA, Kimball AW, (1986), Coronary heart disease mortality and alcohol consumption in Framingham. American Journal of Epidemiology, 124, 481-489.

Fulwood R, Kalsbeek W, Rifkind B, Russell-Briefel R, Muesing R, LaRosa J et al (1986), Total serum cholesterol levels of adults 20-74 years of age. Vital Health Statistics, 11, 236.

Gerhardsson M, Rosenquist U, Ahlbom A and Carlson LA (1986), Serum cholesterol and cancer - a retrospective case-control study. International Journal of Epidemiology, 15, 155-159.

Glass S, Gray M, Eden OB, Hann I, (1987), Scottish validation study of cancer registration data childhood leukemia 1968-81 - 1. Leukemia Research, 11(10), 881-885.

Goldacre M (1986), The Oxford record linkage study: current position and future prospects. In Proceedings of the workshop on computerized record linkage in health research (eds Howe GR, Spasoff RA). University of Toronto Press, pp 97-103.

Goldberg AP (1989), Aerobic and resistive exercise modify risk factors for coronary heart disease. Medical Science in Sports and Exercise, 21, 669-674.

Goldbourt U, Holtzman E and Neufeld HN (1985), Total and high density lipoprotein cholesterol in the serum and risk of mortality: evidence of a threshold effect. British Medical Journal, 290, 1239-1243.

Gordon T, Doyle JT, (1986), Alcohol consumption and its relationship to smoking, weight, blood pressure, and blood lipids. The Albany study. Archives of Internal Medicine, 146, 262-265.

Gordon DJ, Probstfield JL, Garrison RJ, Neaton JD, Castelli WP, Knoke JD, Jacobs DR, Bangdiwala S, Tyroler HA (1989), High-density lipoprotein cholesterol and cardiovascular disease - Four prospective American studies. Circulation, 79, 8-15.

Gordon DJ, Trost DC, Hyde J et al (1987), Seasonal cholesterol cycles: the Lipid Research Clinics Coronary Primary Prevention Trial placebo group. Circulation, 76, 1224-1231.

Harlan WR and Manolio TA (1992), Coronary heart disease in the elderly. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 114-126.

Hartung GH, Foreyt JP, Mitchell RE et al (1983), Effects of alcohol intake on high density lipoprotein cholesterol levels in runners and inactive men. Journal of the American Medical Association, 249, 747-750.

Hartz A, Grubb B, Wild R, van Nort J, Kuhn E, Freedman DS et al (1990), The association of waist hip ratio and angiographically determined coronary artery disease. International Journal of Obesity, 14, 657-665.

Heasman MA (1968), Use of record linkage in long-term prospective studies. In Record linkage in medicine (ed Acheson ED). Edinburgh: Livingstone.

Heasman MA, Clarke JA (1979), Medical Record Linkage in Scotland. Health Bulletin, 37, 97-103.

Hegsted DM and Nicolosi RJ (1987), Individual variation in serum cholesterol levels. Proceedings of the National Academy USA, 84, 6259-6261.

Hole DJ, Clarke JA, Hawthorne VM, Murdoch RM, (1981), Cohort follow-up using computer linkage with routinely collected data. Journal of Chronic Diseases, 34, 291-297.

Holme I (1990), An analysis of randomized trials evaluating the effect of cholesterol reduction on total mortality and coronary heart disease incidence. Circulation, 82, 1916-1924.

Howe GR, Lindsay J (1981), A generalized iterative record linkage computer system for use in medical follow-up studies. Computers and Biomedical research, 14, 327-340.

Hulley SB, Walsh JMB and Newman TB (1992), Health policy on blood cholesterol - time to change directions. Circulation, 86, 1026-1029.

Hunt K, Coleman MP, (1987), The completeness of cancer registration in follow-up studies - a cautionary note. British Journal of Cancer, 56(3), 357-359.

International Collaborative Group (1982), Circulating cholesterol level and risk of death from cancer in men aged 40-69 years. Journal of the American Medical Association, 248, 2853-2859

Inter-society commission for heart disease resources, (1984), Report: optimal resources for primary prevention of atherosclerotic disease. Circulation, 70, 155A-205A.

Isles CG, Walker LM, Beevers GD, Brown I, Cameron HL, Clarke J, Hawthorne V, Hole D, Lever AF, Robertson JWK, Wapshaw JA (1986), Mortality in patients of the Glasgow blood pressure clinic. Journal of Hypertension, 4, 141-156.

Isles CG, Hole DJ, Gillis CR, Hawthorne VM and Lever AF (1989), Plasma cholesterol, coronary heart disease, and cancer in the Renfrew and Paisley survey. British Medical Journal, 298, 920-924.

Iso H, Jacobs DR, Wentworth D, Neaton JD and Cohen JD (1989), Serum cholesterol levels and six-year mortality from stroke in 350,977 men screened for the Multiple Risk Factor Intervention Trial. New England Journal of Medicine, 320, 904-910.

Jabine TB and Scheuren F (1986), Record linkage for statistical purposes: methodological issues. Journal of Official Statistics, 2, 255-277.

Jacobs D, Blackburn H, Higgins M, Reed D, Iso H, McMillan G, Neaton J, Nelson J, Potter J, Rifkind B, Rossouw J, Shekelle R and Yusuf S (1992), Report of the conference on low blood cholesterol: mortality associations. Circulation, 86, 1046-1060.

Jaro MA (1995), Probabilistic linkage of large public health data files. Statistics in Medicine, 14, 491-498.

Jenkins CD, Hames CG, Zyzanski SJ, Rosenman RH, Friedman M, (1969), Psychological traits and serum lipids 1.Findings of the California psychological inventory. Psychosomatic medicine, 31, 115-128.

Kagan A, McGee DL, Yano K, Rhoads GG and Nomura A (1981), Serum cholesterol and mortality in a Japanese-American population. The Honolulu Heart Program. American Journal of Epidemiology, 114, 11-20.

Kannel WB (1983), High density lipoproteins: epidemiologic profile and risks of coronary artery disease. American Journal of Cardiology, 52, 93-123.

Kannel WB (1985), Lipids, diabetes and coronary heart disease: insights from the Framingham study. American Heart Journal, 110, 1100-1107.

Kannel WB (1992), The Framingham experience. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 67-82.

Kannel WB, McGee DL and Castelli WP (1984), Latest perspective on cigarette smoking and cardiovascular disease: the Framingham study. Journal of Cardiac Rehabilitation, 4, 267-277.

Kannel WB, Neaton JD, Wentworth D, Thomas HE, Stamler J, Hulley SB, Kjelsberg MO (1986), Overall and coronary heart disease mortality rates in relation to major risk factors in 325348 men screened for the MRFIT. American Heart Journal, 112, 825-836.

Kannel WB, Wolf PA, Castelli WP and D'Agostino RB (1987), Fibrinogen and risk of cardiovascular disease: the Framingham study. Journal of the American Medical Association., 158, 1183-1186.

Katan MB (1990), Effect of cholesterol lowering treatment on coronary heart disease morbidity and mortality: the evidence from trials, and beyond. Cardiology, 77, suppl. 4, 8-13.

Kendrick S and Clarke J (1993), The Scottish Record Linkage System. Health Bulletin, 51(2), 72-79.

Keys A (1980), In Seven Countries Study: a multivariate analysis of death and coronary heart disease. Harvard University Press.

Keys A, Fidanza F, Karvonen MJ, Kimua N and Taylor HL (1972), Indices of relative weight and obesity. Journal of Chronic Diseases, 25, 329-343.

Kohli HS and Knill-Jones RP (1992), How accurate are SMR1 (Scottish Morbidity Record 1) data? Health Bulletin 50, 14-23.

Kritchevsky SB, Wilcosky TC, Morris DL, Truong KN and Tyroler HA (1991), Changes in plasma lipid and lipoprotein cholesterol and weight prior to the diagnosis of cancer. Cancer Research, 51, 3198-3203.

Kritchevsky SB (1992), Dietary lipids and the low blood cholesterol - cancer association. American Journal of Epidemiology, 135, 509-520.

La Rosa JC (1989), Pravastatin: a new hydrophylic HMG COA reductase inhibitor. In New Advances in the control of lipid metabolism: focus on pravastatin, (ed La Rosa, JC) London: Royal Society of Medicine Services.

Larsson B (1992), Obesity and body fat distribution as predictors of coronary heart disease. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 233-241.

Larsson B, Bjorntorp P and Tibblin G (1981), The health consequences of moderate obesity. International Journal of Obesity, 5, 97-116.

Law MR and Thompson SG (1991), Low serum cholesterol and the risk of cancer: an analysis of the published prospective studies. Cancer Causes and Control, 2, 253-261.

Law MR, Thompson SG and Wald NJ (1994), Assessing possible hazards of reducing serum cholesterol. British Medical Journal, 308, 373-379.

Law MR, Wald NJ and Thompson SG (1994), By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? British Medical Journal, 308, 367-372.

Law MR, Wald NJ, Wu T, Hackshaw A and Bailey A (1994), Sysyematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study. British Medical Journal, 308, 363-366.

Leon AS, Connett J, Jacobs DR, Jr. and Rauramaa R (1987). Leisure-time physical activity levels and risk of coronary heart disease and death. Journal of the American Medical Association, 258, 2388-2395.

Lewis B (1992), Reduction of cholesterol-mediated risk: the role of the doctor. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 343-357.

Lipid Research Clinics Program (1984), The lipid research clinics coronary primary prevention trial results. I. Reduction in incidence of coronary heart disease. Journal of the American Medical Association, 251, 351-364.

Lipid Research Clinics Program (1984), The lipid research clinics coronary primary prevention trial results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. Journal of the American Medical Association, 251, 365-374.

MacFarlane PW, Devine B, Latif S, McLaughlin S, Shoat DB, and Watts MP (1990), Methodology of ECG interpretation in the Glasgow program. Methods of Information in Medicine, 129, 354-361.

MacFarlane PW, Latif S, Shoat DB, and Cobbe SM (1990), Automated serial ECG comparison using the Minnesota code. European Heart Journal, 11, XII European Society of Cardiology Congress abstract supplement, 411.

MacMahon SW, Cutler JD, Furberg CD and Payne GH (1986), The effects of drug treatment of hypertension on morbidity and mortality from cardiovascular disease: a review of randomised trials. Progress in Cardiovascular Disease, suppl. 29, 99-118.

Manninen V, Elo O, Frick MH, Haapa K, Heinonen O, Heinsalmi P, Helo P, Huttenen JK, Kaitaniemi P, Koskinen P, Maenpaa H, Malkonen M, Manttari M, Norola S, Pasternack A, Pikkarainen J, Romo M, Sjoblom T, Nikkila EA (1988), Lipid alterations and decline in the incidence of coronary heart disease in the Helsinki Heart Study. Journal of the American Medical Association, 260, 641-651.

Manninen V, Tenkanen L, Koskinen P, Huttenen JK, Manttari M, Heinonen OP, Frick MH (1992), Joint effects of serum triglyceride and LDL cholesterol and HDL cholesterol concentrations on coronary heart disease risk in the Helsinki Heart Study. Implications for treatment. Circulation, 85, 37-45.

Manson JE, Stempfer MJ, Hennekens CH and Willett WC (1987), Body weight and longevity, a reassessment. Journal of the American Medical Association, 257(3), 353-358.

Manson JE, Colditz G, Stempfer MJ, Willett WC, Rosner B, Monson RR et al (1990), A prospective study of obesity and risk of coronary heart disease in women. New England Journal of Medicine, 322,882-889.

Marenah CB, Lewis B, Hassall D, LaVille A, Cortese C, Mitchell WD, Bruckdorfer KR, Slavin B, Miller NE, Turner PR and Heduan E, (1983), Hypocholesterolaemia and non-cardiovascular disease: metabolic studies on subjects with low plasma cholesterol concentrations. British Medical Journal, 286, 1603-1606.

Marmot M (1992), Coronary heart disease: rise and fall of a modern epidemic. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 3-19.

Marmot M (1994), The cholesterol papers - Lowering population cholesterol concentrations probably isn't harmful. British Medical Journal, 308, 351-352.

Marmot MG, Rose G, Shipley MJ and Thomas BJ (1981), Alcohol and mortality: a U-shaped curve. Lancet, 1, 580-583.

Marmot MG, Shipley MJ and Rose G (1984), Inequalities in death - specific explanations of a general pattern? Lancet, 1, 1003-1006.

Martin MJ, Hulley SB, Browner WS, Kuller LH (1986), Serum cholesterol, blood pressure, and mortality: implications from a cohort of 361662 men. Lancet, 2, 933-936.

Mcloone P (1994), Carstairs scores for Scottish postcode sectors from the 1991 census. Glasgow: Public Health Research Unit, University of Glasgow.

McLoone P and Boddy FA (1994), Deprivation and mortality in Scotland, 1981 and 1991, British Medical Journal, 309, 1465-1470.

McMichael AJ, Jensen OM, Parkin DM and Zaridze DG (1984), Dietary and endogenous cholesterol and human cancer. Epidemiologic Reviews, 6, 192-216.

Meade TW (1987), Epidemiology of atheroma, thrombosis and ischaemic heart disease. In Haemostasis and thrombosis, 2nd edn (eds AL Bloom and DP Thomas), Churchill Livingstone, Edinburgh, pp.697-720.

Meade TW, Mellows S, Brozovic M, Miller GJ, Chakrabarti RR, North WRS et al (1986), Haemostatic function and ischaemic heart disease: principal results of the Northwick Park Heart Study. Lancet, ii, 533-537.

Menotti A, Conti S, Gaimpaoli S, Mariotti S and Signoretti P (1980), Coronary risk factors predicting coronary and other causes of death in fifteen years, Acta Cardiology, 35, 107-120.

Menotti A, Mariotti S, Seccareccia F and Giampaoli S (1987), The 25 year estimated probability of death from some specific causes as a function of twelve risk factors in middle-aged men. European Journal of Epidemiology, 4, 60-67.

Miller SR, Tartter PI, Papatestas AE, Slater G and Aufses AH, (1981), Serum cholesterol and human colon cancer. Journal of the National Cancer Institute, 67, 297-300.

Morris JN (1992), Exercise versus heart attack: history of a hypothesis. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 242-255.

Morris JN, Everitt MG, Pollard R, Chave SPW and Semmence AM (1980), Vigorous exercise in leisure-time: protection against coronary heart disease. Lancet, ii, 1207-1210.

Muldoon MF, Manuck SB, Matthews KA (1990), Lowering cholesterol concentrations and mortality: a quantitative review of primary prevention trials. British Medical Journal, 301, 309-313.

Multiple Risk Factor Intervention Trial Research Group (1982), Multiple risk factor intervention trial: risk factor changes and mortality results. Journal of the American Medical Association, 248, 1465-1477.

Neutel CI, Johansen HL and Walop W (1991), 'New data from old': epidemiology and record linkage. Progress in Food and Nutrition Science, 15, 85-116.

Newcombe HB (1967), Record linking: the design of efficient systems for linking records into individual and family histories. American Journal of Human Genetics, 19, 335-359.

Newcombe HB (1988), Handbook of record linkage. Oxford: Oxford University Press.

Newcombe HB, Kennedy JM, Axford SJ and James AP (1959), Automatic linkage of vital records. Science, 130, 954-959.

Newcombe HB, Smith ME, Howe GR, Mingay J, Strugnell A and Abbatt JD (1983), Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. Computers in biology and medicine, 13, 157-169.

Newcombe HB, Fair ME, LalondeP (1992), The use of names for linking personal records. Journal of the American Statistical Association, 420, 1193-120.

Oliver MF, (1992), Is cholesterol reduction always safe? European Journal of Clinical Investigation, 22, 441-442.

Parish S, Collins R, Peto R, Youngman L, Barton J, Jayne K, Clarke R, Appleby P, Lyon V, Cederholm-Williams S, Marshall J and Sleight P, for the International Studies of Infarct Survival (ISIS) Collaborators (1995), Cigarette smoking, tar yields, and non-fatal myocardial infarction: 14000 cases and 32000 controls in the United Kingdom. British Medical Journal, 311, 471-477.

Pekkanen J, Nissinen A, Punsar S, Karvonen MJ, (1989), Serum cholesterol and risk of accidental or violent death in a 25 year follow-up. The Finnish cohorts of the Seven Countries Study. Archives of Internal Medicine, 149, 1589-1591.

Peterson B, Kristenson H, Sternby NH, Trell E, (1980), Alcohol consumption and premature death in middle-aged men. British Medical Journal, 1, 1403-1406.

Pocock SJ, Shaper AG, Phillips AN, Walker M, Whitehead TP (1986), High-density lipoprotein is not a major risk factor for ischaemic heart disease in British men. British Medical Journal, 292, 515-519.

Pocock SJ, Cook DG, Shaper AG, Phillips AN and Walker M (1987), Social class differences in ischaemic heart disease in British men. Lancet, 2, 197-201.

Pocock SJ, Shaper AG, Phillips AN (1989), Concentrations of high density lipoprotein cholesterol, triglycerides, and total cholesterol in ischaemic heart disease. British Medical Journal, 298, 998-1002.

Pooling Project Research Group (1978), Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report. Journal of Chronic Diseases, 31, 201-306.

Poulter N (1991), Management of multiple risk factors for coronary heart disease in patients with hypertension. American Heart Journal, 121, 246-249.

Pregibon D (1981), Logistic regression diagnostics. Annals of Statistics, 9m 705-724.

Registrar General Scotland (1992), Annual Report for 1991. Edinburgh: General Register Office.

Registrar General Scotland (1994), Annual Report for 1993. Edinburgh: General Register Office.

Rissanen A, Heliovaara M, Knekt P, Aromaa A, Reunanen A and Maatela J (1989), Weight and mortality in Finnish men, Journal of Clinical Epidemiology, 14, 32-38.

Rose G and Shipley MJ (1980), Plasma lipids and mortality: a source of error. Lancet, 523-526.

Ryman A (1994), Cholesterol, violent death, and mental disorder. The association deserves further specific study. British Medical Journal, 309, 421-422.

Salmond CE, Beaglehole R and Prior IAM (1985), Are low cholesterol values associated with excess mortality? British Medical Journal, 290, 422-424.

SAS Institute Inc. (1993), SAS Technical Report P-243, SAS/STAT Software: The GENMOD Procedure, release 6.09; Cary, NC.

Scandinavian Simvastatin Survival Study Group (1994), Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The Scandinavian Simvastatin Survival Study (4S). Lancet, 344, 1383-1389.

Schatzkin A, Hoover RN, Taylor PR, Ziegler RG, Carter CL, Albanes D, Larson DB and Licitra LM (1988), Site-specific analysis of total serum cholesterol and incident cancer in the National Health and Nutrition Examination Survey 1 Epidemiologic Follow-up Study. Cancer Research, 48, 452-458.

Schouten LJ, Hoppener P, Van den Brandt PA, Knottnerus JA and Jager JJ (1993), Completeness of cancer registration in Limburg, the Netherlands. International Journal of Epidemiology, 22, 369-376.

Scottish Record Linkage Team (1995), 'One pass' and 'best-link' methods: enhancing the efficiency of probability-based record linkage. Information and Statistics Division Working Paper.

Shannon HS, Jamieson E, Walsh C, Julian JA, Fair ME, Buffet A, (1989), Comparison of individual follow-up and computerized record linkage using the Canadian mortality database. Canadian Journal of Public Health, 80, 54-57.

Shaper AG and Elford J (1992), Regional variaitions in coronary heart disease in Great Britain: risk factors and changes in environment. In Coronary heart disease epidemiology, (eds Marmot M and Elliott P), Oxford: Oxford University Press, pp. 127-139.

Shepherd J, Betteridge DJ, Durrington P, Laker M, Lewis B, Mann J, Reckless JPD, Thompson GR (1987), Strategies for reducing coronary heart disease and desirable limits for blood lipid concentrations: guidelines of the British Hyperlipidaemia Association. British Medical Journal, 295, 1245-1246.

Sherwin RW, Wentworth DN, Cutler JA, Hulley SB, Kuller LH and Stamler J (1987), Serum cholesterol levels and cancer mortality in 361662 men screened for the Multiple Risk Factor Intervention Trial. Journal of the American Medical Association, 257, 943-948.

Silberberg JA, Henry DA (1991), The benefits of reducing cholesterol levels: the need to distinguish primary from secondary prevention. 1. A meta-analysis of cholesterol-lowering trials. Medical Journal of Australia, 155, 665-670.

Simmons H (1989), Record matching and linking: a technical guide. Oxford Unit of Clinical Epidemiology (Draft copy).

Simons LA (1986), Interrelations of lipids and lipoproteins with coronary artery disease mortality in 19 countries. American Journal of Cardiology, 57, suppl., 5G-10G.

Sletten IW, Nilsen JA, Young RC, Anderson JT, (1964), Blood lipids and behaviour in mental hospital patients. Psychosomatic medicine, 26, 261-266.

Smith WCS, Crombie IK, Tavendale R, Irving JM, Kenicer MB, Tunstall-Pedoe H, (1987), The Scottish Heart Health Study : Objectives and development of methods. Health Bulletin, 45, 211-217.

Smith WCS, Tunstall-Pedoe H, Crombie IK, Tavendale R, (1989), Concomitants of excess coronary deaths - major risk factor and lifestyle findings from 10359 men and women in the Scottish heart Health Study. Scottish Medical Journal, 34, 550-555.

Smith WCS, Shewry MC, Tunstall-Pedoe H, Crombie IK and Tavendale R (1990), Cardiovascular disease in Edinburgh and north Glasgow - a tale of two cities. Journal of Clinical Epidemiology, 43, 637-643.

Sorlie PD and Feinleib M (1982), The serum cholesterol - cancer relationship: an analysis of time trends in the Framinghan Study. Journal of the National Cancer Institute, 69, 989-996.

St Leger AS, Cochrane AL and Moore F (1979), Factors associated with cardiac mortality in developed countries with particular reference to the consumption of wine. Lancet, 1, 1017-1020.

Stamler J, Neaton JD and Wentworth DN (1989), Blood pressure (systolic and diastolic) and risk of fatal coronary heart disease. Hypertension, 13, 2-12.

Stocks P (1944), Measurement of morbidity. Proceedings of the Royal Society of Medicine, 37, 593-608.

Stokes J, Kannel WB, Wolf PA, Cupples LA, D'Agostino RB (1987), The relative importance of selected risk factors for various manifestations of cardiovascular disease among men and women from 35-64 years old: 30 years of follow-up in the Framingham Study. Circulation, 75, suppl. V, 65-73.

Sytkowski PA, Kannel WB, D'Agostino RB (1990), Changes in risk factors and the decline in mortality from cardiovascular disease. The Framingham Heart Study. New England Journal of Medicine, 322, 1635-1641.

Tepping BJ (1968), A model for optimum linkage of records. Journal of the American Statistical Association, 63, 1321-1332.

Thelle D, (ed), (1991). Multiple risk factors for coronary heart disease: identificaton and intervention. Merck, Sharp and Dohme.

Thompson SG and Pocock SJ (1990), The variability of serum cholesterol measurements: implications for screening and monitoring. Journal of Clinical Epidemiology, 43, 783-789.

Tornberg SA, Holm LE, Carstensen JM and Eklund GA (1989), Cancer incidence and cancer mortality in relation to serum cholesterol. Journal of the National Cancer Institute, 81, 1917-1921.

Tunstall-Pedoe H, Smith WCS, Crombie IK, (1986), Level and trends of coronary heart disease mortality in Scotland compared with other countries. Health Bulletin, 44, 153-161.

Tunstall-Pedoe H, Smith WCS, Crombie IK, Tavendale R, (1989), Coronary risk factor and lifestyle variation across Scotland: results from the Scottish Heart Health Study. Scottish Medical Journal, 34, 556-560.

Vartiainen E, Puska P, Pekkanen J, Tuomilehto J, Lonnqvist J and Ehnholm C (1994), Serum cholesterol concentration and mortlaity from accidents, suicide, and other violent causes. British Medical Journal, 309, 445-447.

Virkkunen M (1983), Serum cholesterol levels in homicidal offenders. A low cholesterol level is connected with a habitually violent tendency under the influence of alcohol. Neuropsychobiology, 10, 65-59.

Virkkunen M and Penttinen H (1984), Serum cholesterol in aggressive conduct disorder: a preliminary study. Biological Psychiatry, 19, 435-439.

Waaler HT (1983), Height, weight, and mortality: the Norwegian experience. Acta. Med. Scand., Suppl 3, 679,

Wallace RB, Rost C, Burmeister LF and Pomrehn PR (1982), Cancer incidence in humans: relationship to plasma lipids and relative weight. Journal of the National Cancer Institute, 68, 915-918.

West of Scotland Coronary Prevention Study Group (1992), A coronary primary prevention study of Scottish men aged 45-64 years: trial design. Journal of Clinical Epidemiology, 45, 849-860.

West of Scotland Coronary Prevention Study Group (1995), Screening experience and baseline characteristics in the West of Scotland Coronary Prevention Study. American Journal of Cardiology, 76, 485-491.

West of Scotland Coronary Prevention Study (In press), Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. Journal of Clinical Epidemiology, in press.

WHO MONICA Project Principal Investigators, (1988), The World Health Organisation MONICA Project (Monitoring trends and determinants in cardiovascular disease) : a major international collaboration. Journal of Clinical Epidemiology, 41, 105-114.

Wilkes HC, Kelleher C and Meade TW (1988), Smoking and plasma fibrinogen. Lancet, 1, 307-308.

Williams, RR., Sorlie, PD, Feinleib M, McNamara PM, Kannel WB and Dawber TR (1981), Cancer incidence by levels of cholesterol. Journal of the American Medical Association, 245, 247-252.

Winkler WE (1989), Methods for adjusting for lack of independence in an application of the Fellegi-Sunter model of record linkage. Survey Methodology, 15, 101-117.

Wysowski DK and Gross TP (1990), Deaths due to accidents and violence in two recent trials of cholesterol lowering drugs. Archives of Internal Medicine, 150, 2169-2172.

Yusuf S, Wittes J and Friedman L (1988), Overview of results of randomised clinical trials in heart disease II. Unstable angina, heart failure, primary prevention with aspirin and risk factor modification. Journal of the American Medical Association, 260, 2259-2263.