

Chapter 2: Methods

Box 2: Overview of Chapter 2

Structure of this chapter

This chapter is structured systematically according to the key tasks of conducting a cohort study:

- Defining the study exposures (in this case socioeconomic factors) and covariates – Section 1.
- Defining the study outcomes (mortality) – Section 2.
- Following-up the cohort over time for the outcome of interest (ie, the record linkage of census and mortality records) – Section 3.
- Analysis of the cohort study data – Section 4.

Exposures and covariates

The study-base was the 1991 census data set. Socioeconomic 'exposures' derived from this data set included small area deprivation, education, labour force status, car access, housing tenure and household income. Covariates included age, sex, ethnicity, receipt of a sickness benefit and marital status.

Outcome

Mortality in the three years following census night, for people aged 0–74 years on census night.

Follow-up

The cohort was assembled by anonymously and probabilistically linking 1991 census records to mortality records for 1991–94. Automatch® software was used to conduct the record linkage.

Analysis

First, mortality records that were linked to a census record were compared to those that were not linked to determine the bias in the record linkage by demographic and socioeconomic factors (ie, linkage bias). Stratified analysis and log-linear regression methods were used to measure this bias.

Second, the association of socioeconomic factors measured on the 1991 census (education, small area deprivation, occupational class, housing tenure, car access, and income) with all- and cause-specific mortality were determined. These associations were estimated using logistic regression models within four sub-populations: 25–44 year old males; 45–64 year old males; 25–64 year old females; and 45–64 year old females. Numerous sensitivity analyses were conducted to determine the likely impacts of selection bias, health selection, and linkage bias. Multivariate analyses explored the direct and indirect associations of education, car access, income and labour force status with mortality.

1 Socioeconomic exposures and covariates

The 1991 New Zealand Population Census was used as the study-base. All the independent variables were derived from this census data.

A Census of Population and Dwellings occurs every five years in New Zealand, as mandated under Section 23 of the Statistics Act 1975. Each census undergoes extensive questionnaire development, pre-testing and pilot surveys (Department of Statistics 1992a).

The census in New Zealand has a high response rate – although inevitably some people are missed. The actual undercount and how it varies by demographic groups for the 1991 census (and all previous censuses) is unknown. However, a Post Enumeration Survey (PES) was conducted following the 1996 census (Ewing 1997). It is likely that the results from the 1996 Post Enumeration Survey would approximately apply to the 1991 census. The 1996 PES was a random sample of 25,000 people interviewed two weeks after census night. As the PES ‘used more tightly controlled collection procedures, and more highly trained and experienced field staff’ than the 1996 census itself, the PES was assumed to be a valid tool to estimate the 1996 census undercount (Ewing 1997). Results from the 1996 PES estimated that 1.2% of the New Zealand population were not counted on census night. The estimated undercount also varied by demographic groups:

- 1.4% for males and 1.0% for females
- 1.4%, 2.1%, 0.9%, and 0.6% for 0–14, 15–29, 30–44 and 45+ year olds, respectively
- 2.9%, 3.1%, and 0.8% for Maori, Pacific peoples, and New Zealand Europeans, respectively
- 1.3%, 1.0%, and 0.8% for northern North Island, rest of North Island, and South Island, respectively.

As these undercounts are all small, they are unlikely to cause the census data to be unrepresentative of the total New Zealand population. The age group included for analysis in this report (25–64 year olds) had a less than 1% overall undercount. These undercounts may, however, cause some variation in record linkage success between demographic groups – a small percentage of decedents in 1991 to 1994 would not have completed the 1991 census, and thus fail to have their mortality record linked to their census record.

The general groupings of questions in the 1991 census were:

- activities and voluntary work
- demographics (name and address, sex, age, ethnicity (self-identified and Maori ancestry), country of birth, marital status)
- education
- employment and labour force status
- income and income support
- residence (usual, night of census, five years ago)
- means of travel to work, industry, occupation, and name of employer
- relationship to occupier and living arrangements.

The relevant individual-level socioeconomic variables that could be derived from the 1991 census questionnaire were:

- education
- labour force status
- occupational class

- housing tenure
- car access
- household income.

In addition, small area deprivation was also derived from meshblock geocodes on the census data set. These socioeconomic factors are presented in more detail under subsequent subheadings, followed by the main covariates (sex, age, and ethnic group) and other indicator variables.

1.1 Small area deprivation

The NZDep91 (1991 New Zealand small area deprivation index (Salmond et al 1998)) was developed using 1991 census data. Approximately 20,000 small areas were formed from about 35,000 meshblocks – meshblocks are Statistics New Zealand’s (SNZ) smallest geographic unit, with a median population of about 90. Ten variables (proportions of people/households in the small area) that reflected a lack of something were used to create the index, and reflected seven dimensions of deprivation: income, transport, living space, home ownership, employment, qualifications and support. The factor structure of the first two principal components suggested a single underlying construct of deprivation. The resultant NZDep91 index is the weighted (weights in the first principal component) sum of the ten standardised variables for each small area. The distribution of NZDep91 scores is highly skewed – there is much discrimination among the more deprived small areas, but little discrimination among the least deprived small areas.

Small area deprivation was not directly elicited by the census questionnaire, but by later assigning the NZDep91 score to census records by use of the usual residence meshblock code. Decile and quintile categories of NZDep91 were used in the analysis.

Small area deprivation mortality gradients have already been described in New Zealand (Salmond and Crampton 2000). The purpose of analysing deprivation mortality gradients in the NZCMS was therefore not to report new findings, but to:

- measure bias in the record linkage
- measure possible selection biases in the cohort analyses when the cohort has to be restricted to just those records with complete information
- measure possible health selection effects in the cohort analyses when the cohort has to be restricted to just those respondents in the labour force on census night.

1.2 Education

Educational attainment was elicited with two questions on the 1991 census personal questionnaire. The first (Question 16) was ‘What is your highest school qualification?’, with seven mutually exclusive answers:

- no school qualification
- School Certificate in one or more subjects
- Sixth Form Certificate or University Entrance in one or more subjects
- Higher School Certificate or Higher Leaving Certificate
- University Bursary or Scholarship
- overseas qualification (such as United Kingdom GCE)
- other school qualification (please state).

The second question (Question 17) was ‘What educational or job qualifications have you obtained since leaving school?’, with 11 answer options (multiple selections permitted):

- no qualifications since leaving school
- still at school
- trade certificate or advanced trade certificate
- nursing certificate or diploma
- New Zealand certificate or diploma
- technicians certificate
- teachers certificate or diploma
- university certificate or diploma below Bachelor level
- Bachelors Degree
- postgraduate degree, certificate or diploma
- other qualifications (such as ACA, local polytechnic certificate or diploma) (please state).

From these two questions SNZ derived three separate measures of educational attainment: highest school qualification, tertiary qualifications, and highest gained qualification (ie, highest school or tertiary). Note that the census questions allow a measure of educational attainment, not academic ability or performance.

As many people had no tertiary qualification, the variable ‘tertiary qualifications’ was not suitable for ranking individuals by education in the NZCMS. A selection between highest school qualification and highest gained qualification at both school and tertiary institutes was required. Highest school qualification had some initial appeal in that everyone is exposed to a similar schooling experience. However, a substantial number of people had an overseas school qualification that was unable to be ranked compared to the New Zealand school qualifications. Moreover, while many of the ‘other’ school qualifications would have been equivalent to school certificate, or alternative technical qualifications, the ranking was not clear. An empirical advantage of the ‘highest gained qualification’ variable regarding these two categories was that many people with ‘other’ or ‘overseas’ school qualifications obtained a higher tertiary qualification that was more readily ranked.

A further disadvantage of ‘highest school qualification’, compared to ‘highest gained qualification’, was that the group with no school qualifications was large (45.7% of 25–64 year old males, and 42.4% of 25–64 year old females). This group with no school qualifications was also heterogeneous with regard to subsequent technical and trade training. Thus, the ‘highest gained education’ variable categorised many of those individuals with no school qualification to a trade or technical tertiary qualification, reducing the size of the groups with no qualifications for the ‘highest gained education’ variable (32.3% of 25–64 year old males, and 36.6% of 25–64 year old females).

Finally, the ‘highest gained education’ variable was more comparable to educational attainment variables used in international research.

Table 5: Categories of the highest gained education variable used in the NZCMS

Categories	%	Description
1 Graduate and postgraduate	7.9	Postgraduate degrees (eg, Masters and PhD) and diplomas, and Bachelors degrees.
2 Undergraduate, technical, and teaching	13.4	Undergraduate certificates and diplomas, technician certificates, teachers certificate and diploma, nursing certificate and diploma.
3 Trade certificates, other tertiary	21.4	Trade certificates, other tertiary.
4 11–12 years of school	8.1	University bursary, university scholarship, higher-school leaving certificate, sixth form certificate, and university entrance.
5 10 years of school	12.0	School certificate.
6 Other school qualification	2.7	Other New Zealand school qualification, and overseas school qualification.
7 Nil	34.5	No qualifications.

Note: Percentages are of the combined 25–64 year old male and female population at their usual residence on census night, where the dwelling was a private dwelling population.

The categories of this educational variable are shown in Table 5. The ordering of the variables is approximately from highest to lowest educational attainment. An underlying assumption in the ranking was that any post-school qualification was higher than any school-based qualification.

1.3 Labour force status

Labour force status was elicited with a sequence of questions on the personal questionnaire. The first screening question (Question 21) was ‘Do you work in a job, business, farm or profession?’, with answers of yes or no. Those answering no were then directed to three further questions:

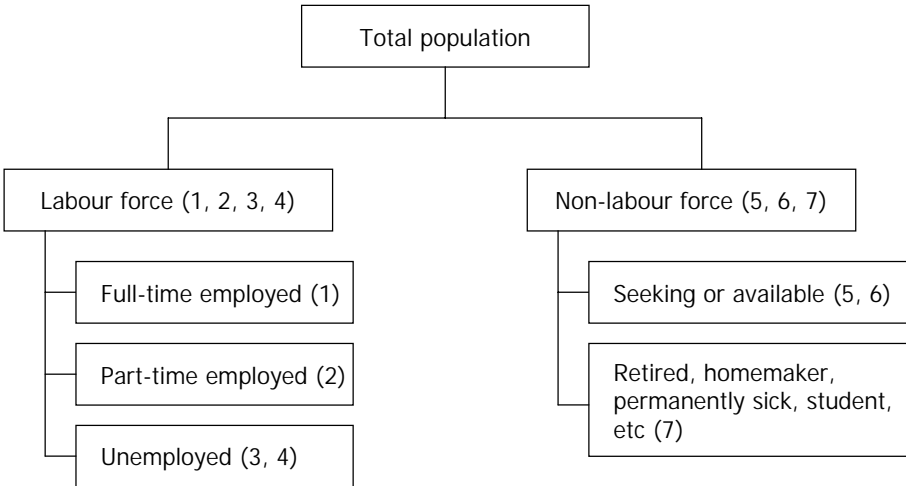
- ‘Did you look for paid work in the last four weeks?’, with three possible answers:
 - no
 - yes – looked for full-time work (full-time work is 30 or more hours per week)
 - yes – looked for part-time work (part-time work is less than 30 hours per week).
- ‘What methods did you use to look for paid work?’, with five possible answers (multiple answers allowed):
 - looked at job advertisements in newspapers
 - contacted the Department of Labour’s New Zealand Employment Service
 - wrote, phoned or applied in person to an employer
 - contacted friends or relatives for help in finding a job
 - other methods (such as contacted a private employment agency, took steps to set up own business).
- ‘If a job had been available, would you have started last week?’, with answers of yes and no.

On the basis of these three questions, and a question of those in paid work of the number of hours they worked, SNZ assigned all people to one of the seven following categories of labour force status:

- 1 employed full time
- 2 employed part time
- 3 unemployed and actively seeking full-time work
- 4 unemployed and actively seeking part-time work
- 5 not working, seeking work but not available for work
- 6 not working, available for work but not seeking work
- 7 not working, not seeking work nor available for work.

Several levels of aggregation were used for labour force status in the NZCMS (Figure 4). The highest level was simply to dichotomise the population into categories 1 to 4 and categories 5 to 7 listed above. These two categories are considered by internationally accepted definitions to be the 'labour force' and the 'non-labour force' (Department of Statistics 1992a). (This report will usually refer to the former as the '*active* labour force' and the latter as the '*non-active* labour force' to further reduce the possibility for confusion.) *Note that the unemployed were included in the active labour force, not the non-active labour force.* At the second level of aggregation, five labour force status groups were identified: full-time employed, part-time employed, and unemployed within the labour force, and 'seeking or available' and 'retired, homemaker, permanently sick, student, etc' among the non-labour force. The unemployed and 'seeking or available' subsumed categories 3 and 4, and 5 and 6 list above, respectively – they were relatively small categories with little reason to differentiate between them. Third, some of the multivariate cohort analyses specifically sought to assess the role of confounding/mediation of the association of unemployment with particular causes of death (eg, suicide). For these specific analyses, three categories of labour force status were used: the unemployed, employed and non-labour force.

Figure 4: Grouping of labour force status used in the NZCMS



Note: Numbers in parentheses refer to the seven categories listed above for labour force status.

Unfortunately, the New Zealand Census does not specifically determine whether people were out of the labour force for health reasons.

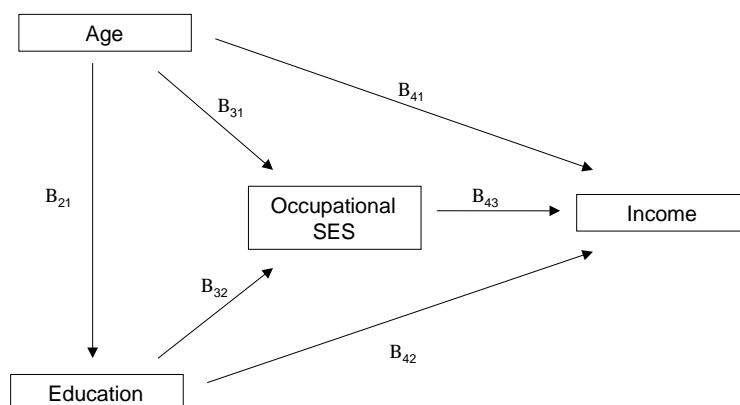
1.4 Occupational class

Information about each individual's occupation was elicited by the questions 'In your main job: (a) What is your occupation? (b) What tasks or duties do you spend the most time on?'. On the basis of the handwritten answers to these questions, coders within SNZ assigned both a NZSCO68 (1968 New Zealand Standard Classification of Occupations) and NZSCO90 (1990) code to each employed individual. (The 1991 census was a transition year from the 1968 to 1990 NZSCO codes, so both codes were assigned.) Occupational class was assigned to each individual using the NZSCO codes with two different measures of occupational status: the New Zealand Socioeconomic Index (NZSEI) and the Elley-Irving scale.

1.4.1 NZSEI occupational class

The NZSEI was developed using 1991 census data (Davis et al 1997; Davis et al 1999b). The NZSEI is premised on the proposition used by Ganzeboom et al (Ganzeboom et al 1992) in their development of the International Socioeconomic Index of Occupational Status, 'that there exists a fundamental relationship between cultural capital or resources (education) and access to material rewards (income), and that this relationship is mediated through the occupational structure' (Davis et al 1997, p.19). Figure 5 below is the path model used in the construction of the NZSEI, with regression coefficients accompanying each arrow. Statistically, one way to fulfil the preceding proposition is to minimise the regression coefficient (B_{42}) directly linking education and income in the path model shown in Figure 5. In effect, this causes the contribution of education (human capital) to income (material rewards) to be channelled as much as possible through occupational socioeconomic position – an indirect causal path. The NZSEI was calculated by an alternating least squares linear regression algorithm that minimised B_{42} . At each iteration, new values of occupational socioeconomic position were assigned to each of the 97 NZSCO90 minor groups. All variables were expressed as standardised continuous variables ('years of education' for the education variable). The output of this process was a scaled occupational socioeconomic position score between 10 and 90 for each of the 97 NZSCO90 minor groups.

Figure 5: Representation of the NZSEI path model



Source: Davis et al, 1997, p.20.

Davis et al (1997) proposed aggregating the NZSCO codes into six occupational classes on the basis of the calculated occupational socioeconomic position score (Table 6 below). Davis et al stated that their division into these six occupational classes was a starting point only – the categorisation has been modified in the NZCMS (Table 6).

Table 6: Alternative classifications of 'occupational class' from NZSEI scores

Class	Range of NZSEI scores (% of 20–69 year old population 1991 census †)				% of 15–64 year old males by Elley-Irving Class, 1986 census ‡
	Davis et al, 1997		NZCMS modification		
1	75–90	5.8%	70–90	10.1%	6.4
2	60–75	17.4%	60–70	13.1%	12.1
3	50–60	20.6%	No change		23.3
4	40–50	22.6%	No change		27.9
5	30–40	16.3%	No change		21.0
6	10–30	17.3%	10–30	8.2%*	9.3
Farmers #	–		22.4, 25.1	9.1%	–

† Derived from Appendix C of Davis et al (1997), giving similar but not identical results to that shown in Table 3.8 of Davis et al (Davis et al 1997).

‡ Taken from Pearce et al 1991 (Pearce et al 1991).

NZSCO90 code 611 (market farmers and crop growers) with an NZSEI score of 22.4; NZSCO90 code 612 (market oriented animal producers) with an NZSEI score of 25.1.

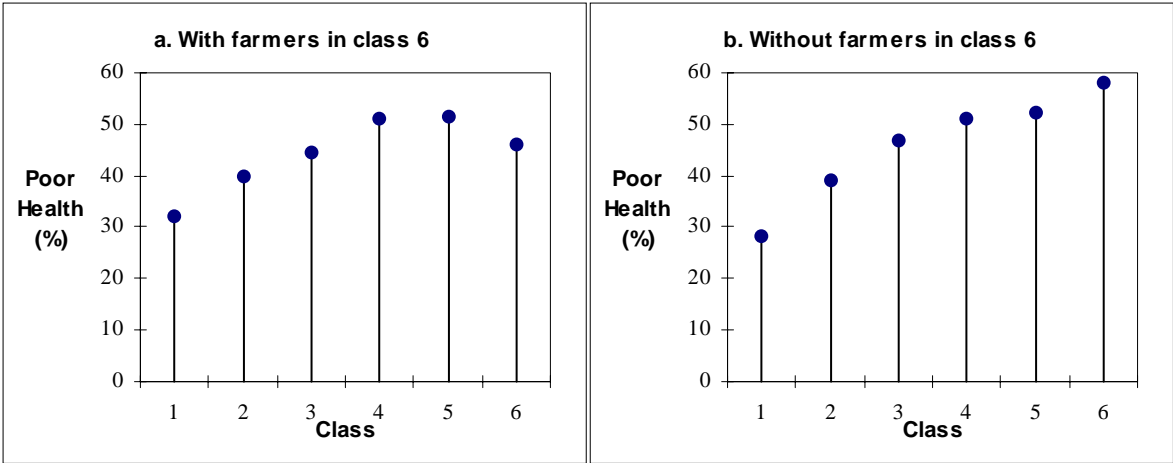
* Excluding farmers.

There were three reasons why the modified NZSEI occupational class classification was preferred in the NZCMS. First, close inspection of NZSCO90 codes allocated by Davis et al to occupational class 2 disclosed a bimodal distribution of NZSEI scores – NZSEI scores were either between 60 and 65, or between 70 and 75. Occupations with a score between 60 and 65 were nursing and midwifery professionals, administrative associate professionals, power generating plant operators, protective service workers, railway engine drivers, primary and early childhood teaching, archivists and librarians, safety and health inspectors, special-interest organisation administrators, physical science and engineering technicians, government associate professionals, and general managers. Occupations with a score between 70 and 75 were business professional, architects and engineers, ship and aircraft controllers, and computing professionals. These latter occupations arguably had more in common for socioeconomic position with the occupations with NZSEI scores above 75 (social and related science professionals, secondary teaching, other teaching professionals, tertiary teaching, life science professionals, physicists and chemists, senior government administrators, mathematicians and statisticians, legislators, legal professionals, senior business administrators and health professionals) than the former occupations with a NZSEI score between 60 and 65.

Second, there are problems with ranking the socioeconomic position of farmers that probably argue for the removal of farmers from the ordinal ranking of occupational classes to be considered as a separate 'special' group. The NZSCO90 classification has just two minor codes for farmers: NZSCO90 code 611 (market farmers and crop growers) with an NZSEI score of 22.4; NZSCO90 code 612 (market-oriented animal producers) with an NZSEI score of 25.1. Within these two groups, there is no distinction between farm owners, farm managers, farm supervisors, and farm workers, and hence a wide distribution of socioeconomic position (Davis et al 1997). Moreover, farm owners are also self-employed, a group known to have a low declared income compared to similar status occupations in New Zealand (Clemance 1985). Occupational class indices used in Europe commonly separate farmers into a separate occupational class (eg, (Kunst et al 1998c)). As generated by the path model used to develop the NZSEI score (Davis et al 1997), the two farming occupation codes both fell within occupational class 6 and 4. Furthermore,

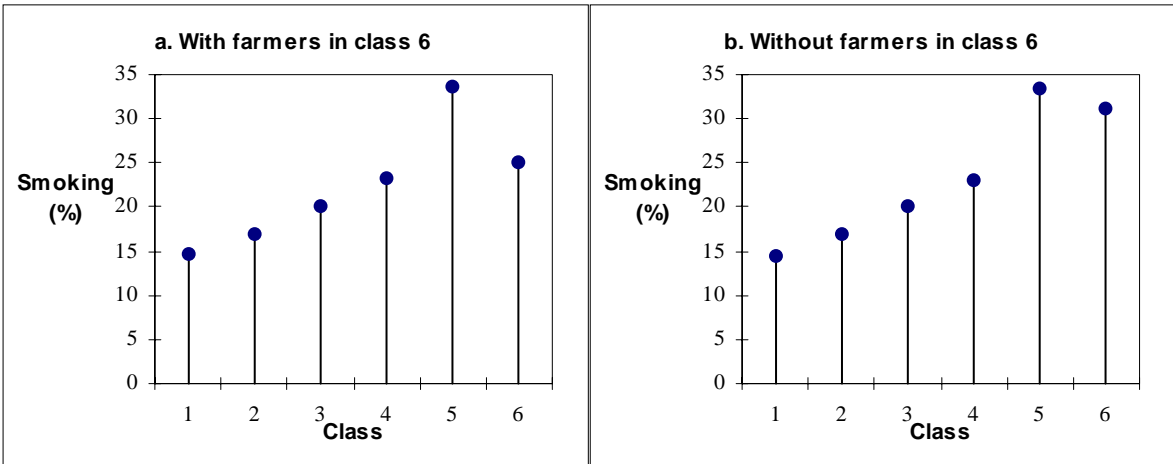
farmers comprised more than half of the proposed occupational class 6 classified by Davis et al. Such ‘misclassification’ is likely to result in underestimation of adverse health effects for occupational class 6: results for poor self-reported health and smoking prevalence in Figure 6 and Figure 7 demonstrate this effect.

Figure 6: Poor self-reported health in the 1992–93 Household Health Survey by NZSEI occupational class, using the classification proposed by Davis et al, but excluding farmers from occupational class 6 in Figure b



Source: Previously unpublished results, kindly forwarded by Keith McLeod, Statistics New Zealand, 1998.

Figure 7: Smoking prevalence in the 1992–93 Household Health Survey by NZSEI occupational class, using the classification proposed by Davis et al, but excluding farmers from occupational class 6 in Figure b



Source: Previously unpublished results, kindly forwarded by Keith McLeod, Statistics New Zealand, 1998.

A similar problem probably also exists for the Armed Forces (Davis et al 1997), but they are relatively few in number (0.6% of employed people), were allocated to a ‘middle’ occupational class (class 3), and so were not separated out. It would be possible to derive NZSEI scores at a lower level of aggregation than the 97 minor occupation groups, that is either for the 260 unit groups, the 563 groups, or some hybrid combination. Such further work may be worthwhile in terms of precision obtained. For example, farmers may be successfully separated and differentiated.

Third, the percentage distribution for the modified classification (excluding farmers) is more symmetric than that proposed by Davis et al, with roughly comparable percentages in occupational class pairs 3 and 4, 2 and 5, and 1 and 6. The percentage of people in occupational class 1 increases from 5.8% to 10.1%, making a more robust comparison group if highest occupational class is used as the reference category. Also, the distribution (excluding farmers) is closer to that for the Elley-Irving scale.

1.4.2 Elley-Irving occupational class

The Elley-Irving socioeconomic index was originally developed using 1966 census data, then revised using 1971 census data (Elley and Irving 1976). The Elley-Irving index per se has not been revised since – the NZSEI should be seen as superseding the Elley-Irving. The Elley-Irving index was developed using education and income, but in a more basic manner than the NZSEI. The median educational attainment (school) and income (personal) were simply determined for each specific occupation, standardised to a common scale, and combined to give a summary score. On the basis of this summary score, six occupational classes were formed.

Davis et al (1997, pp.49–50) compared the Elley-Irving and NZSEI occupational classes, having first derived NZSEI scores for NZSCO68 codes (rather than the NZSCO90 codes). While there was an obvious association of the two indices, there was some overlap in the NZSEI scores between each Elley-Irving occupational class – especially in the middle occupational classes. Davis et al suggest that differences may have arisen due to the different underlying statistical models, and changes in the population between the time the scales were derived. In the NZCMS the NZSEI occupational class was derived from NZSCO90 codes and the Elley-Irving from NZSCO68 codes. As the NZSCO90 codes are a *skills*-based classification, and the NZSCO68 are a *task*-based classification, this generated another potential discrepancy.

1.5 Housing tenure

Housing tenure was measured at the household level, and was elicited on the dwelling questionnaire with the prompt ‘Do the occupants ...’, with options of:

- own this dwelling with a mortgage?
- own this dwelling without a mortgage?
- occupy this dwelling rent free?
- rent or lease this dwelling?

If the last option was selected, the occupier was directed to a further question ‘Who is it rented or leased from?’, with answers of:

- private person
- real estate agency or business organisation
- Housing Corporation
- other government department or corporation, ministry or state owned enterprise
- local authority.

On the basis of these answers, each household in the NZCMS was assigned to one of five categories:

- 1 owned with a mortgage
- 2 owned without a mortgage
- 3 private rental
- 4 public rental
- 5 rental, but not further specified.

Those with free tenancy, or with unspecified tenancy, were assigned as missing values.

1.6 Car access

Car access is measured at the household level, and was elicited with the question 'How many motor vehicles are available for private use by persons in this dwelling?', with answers of none, one, two, three, four, and five or more. The question was on the dwelling questionnaire, completed by the 'occupier'. Occupiers were instructed not to include motor cycles or scooters. In the NZCMS, the answers were aggregated to three categories: none, one, two or more. Unlike income, the number of cars was not equivalised for the number of occupants.

1.7 Equivalised household income

Equivalised total household income was used as the measure of income for each individual in the NZCMS. Thus, each individual in the same household was assigned the same income. The derivation of this variable is justified and described under the following subheadings: household; total household income; equivalisation; and concluding comments.

1.7.1 Household

A household in the 1991 census 'refers to a group of persons, whether related or not, who live together and who normally consume at least one meal together daily or at least share the same cooking facilities' (Department of Statistics 1992b). Households were used as the unit for income in the NZCMS, rather than individuals, as:

'It is within households that income and wealth are pooled, and consumption and savings decisions made. An individual's standard of living is determined not by their income, but by the resources available as a whole to the household in which that individual lives. Thus in examining differences in the standard of living, the natural unit of study is the household.' (Statistics New Zealand 1999, p.51)

In the majority of instances (approximately three-quarters) households were comprised of one family. In the remainder of instances households were single person households, or non-family multi-person households. The use of households, as opposed to families, includes a greater proportion of census respondents.

1.7.2 Total household income

Personal income was determined by question 15 on the 1991 census: 'What will be your total income, including income support, before tax for the year ended 31 March 1991?', with 13 possible options:

- 1 Nil income or loss
- 2 \$2500 or less per year
- 3 \$2501 to \$5000 per year
- 4 \$5001 to \$7500 per year
- 5 \$7501 to \$10,000 per year
- 6 \$10,001 to \$15,000 per year
- 7 \$15,001 to \$20,000 per year
- 8 \$20,001 to \$25,000 per year
- 9 \$25,001 to \$30,000 per year
- 10 \$30,001 to \$40,000 per year
- 11 \$40,001 to \$50,000 per year
- 12 \$50,001 to \$70,000 per year
- 13 \$70,001 and over per year.

Further instructions on the questionnaire were 'Include income from all sources', with prompts of: wages, salary and commission, business or farming income (less expenses), income support, accident compensation weekly payments, interest, dividends, rent, superannuation and pension payments.

A total household income was *unable* to be calculated by SNZ in three circumstances (Department of Statistics 1992a):

- when there were no persons in the dwelling aged 15 years and over
- when there were persons aged 15 years and over absent from their usual residence on census night, *and* the combined income of the persons aged 15 years and over present in the dwelling was less than \$70,000. (If the combined income of the persons present was greater than \$70,000, then regardless of the absent person's/persons' income the household had already met the requirements to be in the top household income category)
- when there were persons aged 15 years and over in the dwelling who did not specify their personal income, *and* the combined income of the remaining persons aged 15 years and over present in the dwelling was less than \$70,000.

In the remaining instances, the total household income was calculated by summing the midpoint income for each individual in the household (ie, \$22,500 if category \$20,001–\$25,000 selected, and \$70,001 if top category selected). This total household income was then recategorised to the same 13 categories as the personal income variable.

1.7.3 Equivalisation

There are economies of scale in households. A family of four does not require four times the income of a single person to enjoy the same standard of living, as, for example, rental costs increase marginally rather than directly proportional to the number of people. Equivalisation is a procedure to adjust household incomes such that they are comparable between households of different size and composition. The revised Jensen Index is a commonly used method to equivalise incomes in New Zealand (Jensen 1988). The Jensen Index assigns equivalences to any combination of adults and children (age less than 18) using the formula:

$$I_{a,c} = (a + wc)^u / 2^u$$

where:

- $I_{a,c}$ = income equivalence of a family household of 'a' adults and 'c' children
- w = the weight for a child compared to an adult
- u = the power parameter to be estimated.

Based on a review of a range of international equivalence scales, and the original Jensen Index (Jensen 1978), Jensen selected two anchor points (values of $I_{a,c}$) for the 1988 revised Index (Jensen 1988): 0.65 for a household of one adult and no children, and 1.75 for a household of two adults and four children. With these two anchor points, values were estimated for w (0.73) and u (0.62). Table 7 shows the thus calculated equivalences for any combination of 1–4 adults and 0–6 children. For example, to have an equivalent standard of living as a two adult household with an income of \$40,000, a household of two adults and two children needs and income of $1.41 \times \$40,000 = \$56,400$.

Table 7: The revised Jensen Index

Number of adults	Number of children						
	0	1	2	3	4	5	6
1	0.65	0.91	1.14	1.34	1.52	1.69	1.85
2	1.00	1.21	1.41	1.58	1.75	1.91	2.06
3	1.29	1.47	1.65	1.81	1.96	2.11	2.25
4	1.54	1.71	1.87	2.02	2.16	2.30	2.44

The revised Jensen Index is not an internationally recognised scale. Perhaps the most commonly used equivalisation procedure internationally (termed the Luxembourg Income Study (LIS) (Atkinson et al 1995) scale here) is simply to divide the household income by the square root of the number of household members (regardless of age). Note that a household income calculated according to the revised Jensen Index is always greater than that calculated according to the LIS scale. For example, a fixed total household income of \$40,000 for a two adult household corresponds to an equivalised household income of \$40,000 by the Jensen Index, but an equivalised household income of only \$28,284 according to the LIS scale. What is important, though, is whether this relative difference between the Jensen Index and LIS scale varies according to household composition. Table 8 shows the ratio of the equivalised household income calculated according to the Jensen Index compared to that according to the LIS scale. Broadly, the relative size of the Jensen to LIS equivalised household income is unchanged with varying numbers of children, but decreases with increasing numbers of adults. Put another way, equivalisation according to the revised Jensen Index 'assumes' less economies of scale for each extra adult in the household than does the LIS scale.

Table 8: Ratio of revised Jensen Index to LIS scale equivalised household income

Number of adults	Number of children						
	0	1	2	3	4	5	6
1	1.54	1.55	1.52	1.50	1.47	1.45	1.43
2	1.41	1.43	1.42	1.41	1.40	1.39	1.38
3	1.35	1.36	1.36	1.36	1.35	1.34	1.34
4	1.30	1.31	1.31	1.31	1.31	1.31	1.30

In the NZCMS, an equivalised household income was calculated by dividing the midpoint of each total household income category by the appropriate value of the revised Jensen Index. For households in the top total household income category (\$70,001 and over per year), the 'mid-point' was taken as \$99,300 based on data from the 1991 Household Economic Survey. When there were more than four adults or more than six children in a household, the value of the Index selected was that for four adults and six children, respectively. When all household members were younger than 18 (but at least one member older than 15), one child was reclassified as an adult.

1.7.4 Concluding comments

As is apparent from the above description, there are a number of possible sources of misclassification. First, the census only uses one question to elicit personal income (rather than multiple questions of salary/wages, interest, dividends, rent, income support, etc), requiring each individual to take a 'best-guess' at their total income. Second, the measure is not disposable income (allowing for taxes and transfers), but gross income. Third, midpoints are assigned to income categories at two stages: to personal income categories to calculate total household income, and to household income to calculate the equivalised household income. Fourth, while '\$70,001 plus per year' in 1991 was a reasonably high top *personal* income category, the same may not be true for *household* income, resulting in some loss of discrimination among high-income households. Finally, there is an inevitable arbitrariness of the selected equivalisation procedure – in this instance, the revised Jensen Index.

1.8 Covariates: sex, age, and ethnicity

1.8.1 Sex

All analyses were conducted separately by sex.

1.8.2 Age

Most analyses were conducted separately for 25–44 year and 45–64 year olds. Within those 20-year age groups, a categorical variable of five-year age groups (eg, 25–29, 30–34, 35–39, and 40–44 years) was included in all analyses.

1.8.3 Ethnicity

The focus of this report was on socioeconomic mortality gradients, not ethnic inequalities in health or the variation in the socioeconomic mortality gradients between ethnic groups. Both are extremely important issues for research, and the NZCMS offers huge potential to unravel some of the overlap between ethnicity and socioeconomic position. Ethnicity and socioeconomic position will be a substantive focus in the NZCMS, but that focus is beyond the scope of this report.

The analyses in this report neither measure ethnic inequalities in mortality in New Zealand, nor measure socioeconomic mortality gradients within ethnic groups.

However, ethnicity is strongly associated with mortality in New Zealand. As the distribution of socioeconomic factors is correlated with ethnicity, and ethnicity is not on the causal chain between socioeconomic position and mortality, ethnicity is a likely confounder of the association of socioeconomic position with mortality (Rothman and Greenland 1998). Accordingly, ethnicity is included as a covariate in all analyses. (Moreover, controlling for ethnicity minimised linkage bias.)

Ethnicity was elicited in the 1991 census with the question ‘Which ethnic group do you belong to?’ (*tick the box or boxes which apply to you*), with answers of: New Zealand European, New Zealand Maori, Samoan, Cook Island Maori, Tongan, Niuean, Chinese, Indian, and Other (please state). SNZ then derived a hierarchical classification:

- New Zealand Maori ethnic group (New Zealand Maori as one of the self-identified ethnic groups)
- Pacific Island Group (any Pacific Island group as one of the self-identified ethnic groups, but not where the individual also self-identified as New Zealand Maori)
- Non-Maori non-Pacific.

In the NZCMS, those not specifying an ethnic group were classified as non-Maori non-Pacific.

1.9 Other indicator variables

In addition to the socioeconomic exposures and covariates of interest, there were a number of other indicator variables that, for example, allowed an estimation of the likely health selection.

1.9.1 Usual residence

This variable identified all the individuals at their usual residence on census night. As the household socioeconomic exposures can only be derived for individuals at their usual residence on census night, most cohort analyses excluded individuals not at their usual residence on census night.

1.9.2 Dwelling type

This variable identified whether the census night dwelling was a private (eg, separate house, caravan, flat) or non-private dwelling (eg, motel, rest home, hospital, prison). Household socioeconomic exposures were also not available for non-private dwellings, so most cohort analyses excluded non-private dwellings.

1.9.3 Sickness beneficiary

Receipt of income support was elicited on the personal questionnaire, with one possible answer being 'Sickness or Invalid's Benefit'. One way to determine the socioeconomic mortality gradient unbiased by health selection is to conduct analyses only upon those individuals that were healthy at the outset of the cohort study. Excluding individuals who received a sickness or invalids benefit allowed a test of possible health selection. However, two opposing biases make this test of uncertain accuracy. First, the sickness benefit is only provided to people below a certain income. This bias would mean that the mortality risk for poor healthy people would be correctly determined, but it would be overestimated for rich healthy people due to residual inclusion of rich sick at the outset. This first bias would cause an underestimate of the socioeconomic mortality gradient. Second, not everyone who is unhealthy (for the purposes of causing a health selection effect) would apply for a sickness benefit. The net effect of these two biases was uncertain.

1.9.4 Hospitalisation before the 1991 census

For the decedents it was possible to determine who had been hospitalised between about 1988 and census night. (Most hospitals began using the NHI number in 1988, allowing death events to be linked to hospital events after that time.) An ideal test of health selection would have been to exclude all members of the *cohort* with a hospitalisation in the three years up to census night, thus allowing analyses on a relatively healthy population at initiation of the cohort study. However, to do so would require linking all people with a hospitalisation event (not just deaths) to the census – an unrealistic exercise. Thus, analysis in the NZCMS could only determine the difference between including all deaths versus including only deaths where the decedent had not been hospitalised in the three years preceding census night.

2 Mortality outcome

Mortality in the three years following census night (5 March 1991) was determined by linking mortality records to census records (see next section for record linkage methods). The mortality records were obtained from New Zealand Health Information Services (NZHIS). In addition to analyses of the association of socioeconomic factors with all-cause mortality, analyses of the association of socioeconomic factors with cause-specific mortality were also conducted. Groupings of cause of death were based on that in the Global Burden on Disease study (Murray and Lopez 1996), with consideration of modifications proposed by Tobias and Christie (1998), as shown in Table 9.

Table 9: ICD codes for groupings of cause-specific deaths used in the NZCMS

Cause of death	ICD codes
Cancer	140–209
Colorectal	153–154
Lung	162
Breast	174
Prostate	185
Cardiovascular disease	410–414, 390–409, 415–459
IHD	410–414
Cerebrovascular	430–438
Infection and pneumonia	001–139, 320–323, 390–392, 460–466, 480–487, 590, 595, 614–616, 680–686, 711, 771
Respiratory	470–478, 490–519
COPD	490–492, 495–496
Unintentional injury	800–949
Road traffic crash	810–825
Other unintentional	800–809, 826–949
Suicide	950–959, 980–989
Homicide, intentional injury	960–979, 990–999
Other	Remaining ICD codes

3 Record linkage

A detailed description of the methods (and results) of the record linkage of mortality and census records can be found in a Technical Report (Blakely et al 1999). Only a summary of the record linkage is included in this chapter, emphasising an epidemiological perspective. In particular, record linkage is presented in terms of a *misclassification bias of the mortality outcome*.

Record linkage has a language of its own. As such, a glossary of terms is provided on page 253. The first use of each term that appears in this glossary is in **bold**.

3.1 Probabilistic record linkage methods

Newcombe (1988) and Baldwin et al (1987) provide good basic introductions to probabilistic record linkage methods (Baldwin et al 1987; Newcombe 1988). The method has been used in the United States to link mortality records to the Current Populations Survey database (Rogot et al 1986). In the last decade, Jaro has developed an advanced software package, Automatch®, for probabilistic record linkage (Jaro 1995; MatchWare Technologies 1998). Automatch® was the software package used in this research.

Humans searching two files for the same individual intuitively do two things. First, they look for agreement or disagreement on variables common on both files (matching variables). Second, they assign varying importance to different variables. For example, a match on a social security number (or some other unique identifier) just about guarantees the records in the two separate files are for the same person. But a match on sex adds only a small amount of discriminatory information. Probabilistic record linkage formalises these intuitive processes, using probability ratios and taking advantage of the processing capacity of computers.

The object of record linkage is to find **matches** between **records** from two (or more) **files**, where a match consists of two records from different files *for the same person*. In the NZCMS, the comparison files were mortality and census records. To achieve this objective, **pairs** of mortality and census records are compared by the variables common to both files – the **matching variables**. It is not possible in any record linkage project to determine exactly which comparison pairs are (correct) matches and non-matches. Rather, pairs are categorised as **links** or **non-links**. It is intended that the majority (hopefully, the vast majority) of links are matches (**true links**), and few matches are falsely assigned as non-links (**false non-links**). A two-by-two table of link/non-link status by match-non-match status is shown in Table 10.

Table 10: Two by two table of link/non-link status by the match/non-match status for comparison pairs in a record linkage project

	Match	Non-match
Link	True links (or true positives)	False links (or false positives)
Non-link	False non-links (or false negatives)	True non-links (or true negatives)

At the heart of probabilistic record linkage are **agreement frequency ratios** and **disagreement frequency ratios**, determined by the ***m* probabilities** and ***u* probabilities**. Consider the variable day of birth (dd), and the value 9. The *m* probability is the probability among the linked records that when dd is 9 for one of the records (eg, mortality), dd is also 9 for the record from the other file (eg, census). The linked records are, to the best of one’s knowledge, correctly matched. But it is neither necessary that each linked record be a correct match, nor necessary that all correct matches be included among the linked records, for the estimation of *m* probabilities to be accurate provided the numbers of false links and false non-links are small. Note that there is a bootstrap problem here – the *m* probability is calculated among the linked records, but before a set of linked records can be created in a probabilistic record linkage process we need to have *m* probabilities. To get around this problem *m* probabilities are initially specified by the operator on the basis of a best guess, and in subsequent iterations of the record linkage the *m* probabilities are updated on the basis of the last set of linked records.

The *u* probability is similar to the *m* probability, except it applies to the non-linked records: the *u* probability is the probability among the non-linked records that when dd is 9 for one of the records (eg, mortality), dd is also 9 for the record from the other file (eg, census). That is, the *u* probability is the probability that for any random comparison pair the given matching variable agrees. This is closely approximated by the frequency of each specific value of each matching variable in the two files.

Having obtained m and u probabilities, agreement and disagreement frequency ratios (or odds) are next calculated for each possible comparison of matching variables. For example, Table 11 gives frequency ratios for agreement and disagreement on dd. Assume that among the linked records, 95% agree on day of birth – the m probability. (The other 5% would disagree principally as a result of coding errors.) Further, assume among the non-linked records 3% agree on day of birth. This 3% is simply estimated by the inverse of the number of possible values of dd, approximately 1/30. That is, among non-links we expect 3% to agree on dd purely by chance. The agreement frequency ratio of 32 to 1 for an observed match on dd is the odds of [the probability of dd agreeing among links] to [the probability of dd agreeing among non-links]. That is, dd is 32 times more likely to agree among links than among non-links. Conversely, the disagreement odds of dd not agreeing among links versus non-links is 1 to 19.

Table 11: Example of agreement and disagreement frequency ratios and weights for comparison by the matching variable 'day of birth'

Comparison outcome	Proportion/frequency		Frequency ratio	Weight
	Links	Non-links		
Agreement	0.95 (m)	0.03 (u)	32/1 (m/u)	4.98 [$\ln(m/u)/\ln(2)$] †
Disagreement	0.05 ($1-m$)	0.97 ($1-u$)	1/19 ($1-m/1-u$)	-4.28 [$\ln(1-m/1-u)/\ln(2)$] †

† The divisor, $\ln(2)$, transforms the natural logarithm to a base 2 logarithm.

Having determined the frequency ratios for each matching variable (and each value of each matching variable), the next step is to calculate the **combined frequency ratio** for any given comparison pair. The combined frequency ratio is the product of the agreement and disagreement frequency ratios for all matching variables, taking the agreement frequency ratio when the matching variable agrees and the disagreement frequency ratio when it disagrees. But the magnitude of the combined frequency ratio quickly becomes very large (for multiple agreements on the matching variables) or very small (for multiple disagreements), and it is easier to use the **combined weight**. The combined weight is the sum of the **agreement** and **disagreement weights**. The agreement or disagreement weight for each matching variable (or value of the matching variable) is the logarithm to base two of the global (or specific) frequency ratio – the formula is given in Table 11. Using logarithms to base two is not necessary, but was the precedent set by researchers involved in pioneering record linkage in the Oxford Record Linkage Study (Baldwin et al 1987). A convenience of using logarithms to base two is that each increase in the weight by one represents a doubling of the overall odds in favour of the comparison pair being a match. The combined weight is then used to allocate, by means of a cut-off, each possible comparison pair to either a set of highly probable pairs (ie, links) or a set of unlikely pairs (ie, non-links).

An additional, and crucial, step in probabilistic record linkage is the **blocking** of records. Blocking involves partitioning the records in both files by a common variable, and then only conducting comparisons of records between files *within these blocks*. For example, two files of 1000 records each could be blocked by age in years, resulting in approximately 10 records in each block in each file. This dramatically reduces the number of comparisons from 1,000,000 without blocking, to $100 \times 10 \times 10 = 10,000$ with blocking by age ([blocks] \times [records in each block in first file] \times [records in each block in second file]). Blocking is thus computationally efficient, and (as described subsequently in the Section 3.3) reduces the number of false links.

3.2 Record linkage in the NZCMS

The record linkage in the NZCMS was anonymous (ie, no names or text address variables were available).

3.2.1 Blocking variables

Five geocodes were extracted from the mortality data (one meshblock code, and four census area units (CAU) codes) that together with the census usual residence meshblock code and CAU code made five possible combinations of blocking variables. These five combinations are shown in Table 12. Meshblocks are the smallest administrative geographic unit in New Zealand, with a median number of 96 people in each meshblock. CAUs are aggregates of meshblocks, containing about 2000 people each.

Table 12: Geocode variables ('blocking' variables) used in the record linkage

Mortality	Census †	Comment
Meshblock (SNZ vitals)	Meshblock	<i>Mortality data.</i> 90.7% of the mortality records were assigned a meshblock code for their usual residence at time of death, by merging the NZHIS mortality records to the SNZ Vitals file. The SNZ Vitals meshblock was derived from the address on the death registration form. <i>Census data.</i> A usual residence meshblock was routinely assigned to all census records by SNZ.
Vitals-CAU	CAU	<i>Mortality data.</i> The CAU containing the SNZ Vitals File meshblock for the above 90.7% of mortality records. While a meshblock was not able to be assigned to the remaining 9.3% of mortality records, the vast majority were directly assigned a CAU. <i>Census data.</i> The CAU containing the usual residence meshblock. This census CAU was used for blocking with each of the four mortality data CAUs.
NHI-CAU	CAU	<i>Mortality data.</i> The NHI file includes an automatically assigned CAU-code on the basis of the text address entered by hospital clerks that maintain the NHI File. Thus it is an independent source of geocode data to the SNZ Vitals.
Post-CAU	CAU	<i>Mortality data.</i> Some decedents were hospitalised after census night, but before the hospitalisation associated with the death event. By linking the NHI and NMDS files at NZHIS, the 'post-CAU' code was derived for the stated usual address at the time of this hospitalisation for some mortality records. The rationale was to obtain a CAU-code for a point in time closer to the census than when the decedent died.
Pre-CAU	CAU	<i>Mortality data.</i> Some decedents were hospitalised between 1988 (when the NHI file was established) and census night. By linking the NHI and NMDS files at NZHIS, the 'pre-CAU' code was derived for the stated usual address at the time of the last hospitalisation before census night for some mortality records. The rationale was to obtain a CAU-code for a point in time closer to the census than when the decedent died.

† Census geocodes are for the usual residence address on census night.

The multiple possible geocodes for the mortality data arose due to the existence of a NHI file (National Health Index file; usually entered by hospital clerks), the NMDS Death Event file (National Minimum Data set; built up from the SNZ-Vitals file and the death registration form), and possibly one or more NMDS Health Event files (hospitalisations; entered by hospital clerks) for each decedent. These NHI and NMDS files could be linked together by NZHIS using the NHI number recorded on all files. The benefit from these multiple and independent sources of data for the mortality records was that CAU codes for the usual residence at time of death *and* points in time closer to census night could be obtained. These multiple CAU codes increased the chance of correctly linking mortality and census records.

Only one meshblock code, however, was available from the mortality data – that on the SNZ Vitals file. NZHIS does not store the meshblock code on the NMDS Death Event file, despite SNZ deriving a meshblock code for the majority of decedents prior to forwarding the data to NZHIS. However, the meshblock derived from the death registration form was easily retrieved. By combining the death registration office, year and number variables a unique identifier was created for each death. Using this unique identifier, the meshblock codes (and Vitals-CAU codes) were transferred from the SNZ Vitals file to the NZHIS mortality data.

3.2.2 Matching variables

The matching variables common to both mortality and census data were sex, ethnic group, country of birth, and date of birth. The latter was disaggregated to day of birth (dd), month of birth (mm), and year of birth (yy). All of these matching variables, except country of birth, were available from both the NMDS Death Event file and the NHI file for each decedent. Thus, the census sex variable could be compared to both the NMDS Death Event file and NHI file sex. The advantage of this double comparison was that if there was a coding error for sex on one of these two mortality files, it was highly unlikely that a coding error would have occurred on both files. These two sources of demographic data for the record linkage therefore increased the discriminatory power of the record linkage process.

3.2.3 Record linkage strategy

The way that these matching and blocking variables were used in the record linkage is described in detail in the Technical Report (Blakely et al 1999). Briefly, the geocodes were used to 'block' the two files, and comparisons of records by the personal 'matching' variables only occurred when the geocodes agreed. Meshblock, as the variable with the greatest number of values, was used as the blocking variable in the first **pass** of the record linkage. Subsequent passes used CAU codes as the blocking variable. In all, eight passes were used in the record linkage.

The final record linkage strategy – in particular the cut-offs and clerical review – was a balance of maximising the number of links obtained (maximising sensitivity), but minimising the estimated percentage of false links (maximising positive predictive value). Three methods were available to estimate the positive predictive value – two specifically developed for the NZCMS. They are complex and not presented in this report – instead, they are described in detail in the Technical Report (Blakely et al 1999, pp.42–59). Briefly, two of the methods were used to estimate the number of false positive links in this report: the chance method and the duplicate method. The chance method is essentially a method for estimating the number of exact links that would occur purely by chance, using the *u* probabilities. The duplicate method uses the number of duplicate pairs (DA pairs; one mortality record linkage to two or more census records) and combinatorial probabilities to estimate the number of false positive links.

For duplicate links, the highest scoring duplicate link was accepted, or if the scores were tied both (or all) links were discarded.

3.3 Probabilistic record linkage as (mis)classification of the mortality outcome, and the effect of blocking

In this section, an example is used to illustrate record linkage in terms of a screening test or tool to ascertain the mortality outcome.

Assume there are 3 million census records and 40,000 deaths in the census cohort in the given follow-up period – the approximate numbers in the NZCMS. Further, assume that the mortality records available for record linkage were these same 40,000 deaths. Thus there was a correct census record match for all 40,000 mortality records somewhere in the census file. Assume that 35,000 of these 40,000 mortality records were correctly linked to their census record – the true links. That leaves 5000 false non-links that were missed due to, for example, coding errors on either file. These numbers and the total of 40,000 matches (or death events) are shown in the first column of Table 13. The linked/non-linked status is the *observed* status of each comparison pair, and the match/non-match status is the *actual* status of each comparison pair.

Table 13: Two by two table of link/non-link status by the match/non-match status in a hypothetical record linkage example *without* blocking

	Match	Non-match	
Link	35,000 (true links)	1,200,000 (false links)	1,235,000
Non-link	5000 (false non-links)	1.1999876×10^{11} (true non-links)	$1.19998765 \times 10^{11}$
	40,000	1.1999996×10^{11}	1.2×10^{11}

The total number of possible comparison pairs for these 40,000 mortality records and 3 million census records is $40,000 \times 3,000,000 = 1.2 \times 10^{11}$. Subtracting the 40,000 matches leaves 1.1999996×10^{11} non-matches. Assume that the matching variables were as in the NZCMS – dd, mm, yy, sex, ethnicity, country of birth – and that (for simplicity in this example) exact agreement was required on each of these variables. (Probabilistic record linkage allows for some disagreement on some matching variables. See the Technical Report for details of how this was specified in the NZCMS (Blakely et al 1999).) The probability of any randomly selected non-match pair agreeing on all these variables is the product of the u probabilities, in this case approximately $1/30 \times 1/12 \times 1/60 \times 1/2 \times 2/3 \times 2/3 = 0.00001$. (Here, $1/60$ is an approximate ‘average’ u probability for yy, and $2/3$ is an approximate ‘average’ u probability for ethnicity and country of birth.) Thus, $0.00001 \times 1.1999996 \times 10^{11} = 1,199,999.6 \approx 1.2$ million of the non-matches would be categorised as links – false links in the top right cell of the two-by-two Table 13.

In total, therefore, there would be 1,235,000 links in this record linkage example – the top row total in Table 6. The percentage of these links that were true links is only 2.8% ($35,000/1,235,000$), an appallingly low positive predictive value in screening terms that will cause ruinous bias in any cohort study! However, **blocking** can substantially improve the positive predictive value. Assume that there exists a geocode with 30,000 values that blocks the census and mortality files into blocks containing, on average, 100 census records and 1.33 mortality records in each block (ie, similar to the meshblock used in the NZCMS). With any blocking strategy, the cost is that if the true matches disagree on the blocking variable, that match is missed in the record linkage and becomes a false non-link – a problem known as ‘skipping’. Assume in our example that skipping reduces the number of true links to 30,000 as shown in Table 14. Thus the introduction of a blocking variable reduces the sensitivity of the record linkage in this example from 87.5% ($35,000/40,000$) to 75% ($30,000/40,000$).

Table 14: Two by two table of link/non-link status by the match/non-match status in a hypothetical record linkage example *with* blocking

	Match	Non-match	
Link	30,000 (true links)	40 (false links)	30,040
Non-link	10,000 (false non-links)	3,989,960 (true non-links)	3,999,960
	40,000	3,990,000	4,030,000

In addition to the gain in computing efficiency brought about by blocking due to fewer necessary comparisons, there is also a substantive gain in the PPV. In this example, there are now $30,000 \times 1.33 \times 100 = 3,990,000$ possible comparison pairs (ie, [number of blocks] \times [average number of mortality records in each block] \times [average number of census records in each block]). As above, 0.00001 of these comparison pairs will be categorised as links due to a purely chance agreement on the matching variables, ie, $0.00001 \times 3,990,000 = 39.9 \approx 40$. The PPV now in this example is a very respectable 99.9% ($30,000/30,040$). The cost of obtaining this improvement in PPV was a drop in sensitivity. Note that the specificity remained unchanged between Tables 13 and 14 ($1.1999796 \times 10^{11}/1.1999996 \times 10^{11} = 3,989,960/3,990,000 = 0.9999$).

This above example is simplistic. For example, in Table 13 there are only 40,000 mortality records yet over 2 million links (ie, each mortality record is linked to 50 census records on average). Further, the above example does not allow for varying block sizes, duplicate links (ie, one mortality (census) record linked to two or more census (mortality) records), and partial disagreements on the matching variables. These issues are considered in detail in the Technical Report (Blakely et al 1999). However, the above example demonstrates:

- how the record linkage process is analogous to a screening test for the mortality outcome
- and how blocking increases the PPV by essentially increasing the prevalence of matches in the population of comparison pairs.

4 Data analysis

The data analysis was in two parts: analysis of the linkage bias, and the cohort analysis.

4.1 Linkage bias

In this section, methods are described which are used to quantify the **linkage bias**, where linkage bias is defined as *the biases by demographic and socioeconomic factors in the proportion of mortality records linked to a census record*. (A more detailed description of the methods used in the analysis of linkage bias can be found in the Technical Report (Blakely et al 1999).) A description of linkage bias as a misclassification bias of the mortality outcome is included in Appendix B. It would have been possible to use correction formulas (eg, Copeland et al 1977) to quantify the likely amount of bias of the risk ratios observed in the cohort analyses due to misclassification of the mortality outcome. To do so would first require estimating the sensitivity and specificity of the record linkage for the mortality outcome, including by demographic strata. However, it was simpler, more direct, and probably more accurate, to directly estimate the linkage bias by socioeconomic position (described in the remainder of this section), and then use these estimates to adjust the risk ratios subsequently determined in the cohort analyses. *This adjustment constitutes the sensitivity analyses in the NZCMS of the impact of misclassification bias of the mortality outcome on the observed risk ratios in the cohort analyses.*

Also, note that the linkage bias was determined by comparing the mortality records linked to a census record to those mortality records unlinked to a census record. As such, the mortality records submitted to the record linkage were the total population of analysis. This population approximates, but would not have been exactly the same as, the actual deaths in the census cohort (see Appendix B).

4.1.1 Univariate and stratified analyses

The variation in the proportion of mortality records linked to a census record was determined by demographic factors (sex, age, ethnic group (NHI file), and time between census night and death) and socioeconomic factors (small area deprivation (NZDep91) and occupational class (NZSEI)). Simple categorical methods were used. Estimates of precision (ie, confidence intervals) are not reported for the univariate and stratified analysis – due to the large sample size most differences are statistically significant, and the size of the difference is of greater importance.

4.1.2 Regression analyses

Regression analyses were conducted to determine the ‘independent’ linkage bias due to socioeconomic factors, controlling for demographic factors. The rationale was that the cohort analyses would be conducted within demographic strata (eg, the association of income and mortality among 45–64 year old females – not the association of income and mortality for all people simultaneously). Thus, *the objective was to quantify the residual linkage bias by socioeconomic position within each demographic stratum or group.*

Such analyses could be conducted simply by stratification rather than regression modelling, deriving the ‘actual’ linkage rate by strata. However, this approach was limited as:

- the SNZ protocol is that all absolute cell sizes must be random rounded to a multiple of three. Thus the observed percentage linked in small cells would have been inaccurate for sparse strata
- the assigned ethnic group on the mortality data is not equivalent to that on census data (Blakely et al 2001b) – regression modelling to ‘smooth’ out estimates by strata of ethnic group may be preferable to using actual results from health data ethnic group strata
- census records with either a census night dwelling of ‘private hospital’, ‘public hospital’, or ‘rest-home’, or simply a non-private census night dwelling, were excluded from most of the cohort analyses. But it was not possible to conduct an analysis of bias on a similar restricted set of mortality records as the dwelling type was not recorded in the mortality file. Regression modelling to smooth out estimates by strata may, again, be preferable compared to using actual health data stratum.

Consequently, regression modelling was used to quantify the linkage bias by socioeconomic position. Two different regression-modelling strategies were used. The first strategy attempted to quantify the linkage bias by small area deprivation for all mortality records simultaneously and by occupational class for males and females separately. This is the method described in detail and used extensively in the Technical Report (Blakely et al 1999) – it will only be briefly described here. The second strategy involved simply modelling the linkage bias by small area deprivation and occupational class within each of the four demographic groups used for the cohort analyses in this report – 25–44 and 45–64 year old males, and 25–44 and 45–64 year old females.

The justification of the first modelling strategy was to summarise the linkage bias as much as possible, achieving the most stable estimates possible. However, for the purposes of this report having linkage bias results for each of the four demographic groups allowed direct adjustment for linkage bias in the cohort analyses conducted in this report. Hence, both strategies are described and reported in this report here.

The regression analyses used a generalised linear model with a log link (referred to as log-linear hereafter), conducted in SAS version 6.12. The regression model was:

$$R(x) = \exp(\alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n)$$

where:

$R(x)$ is the average risk of being linked to census record given covariates x

x_1, x_2, \dots, x_n are the covariates or interaction products (eg, sex, age group)

$\beta_1, \beta_2, \dots, \beta_n$ are the coefficients

α is the intercept.

The log-linear risk models were fitted with a binomial error term, Pearson estimation methods, and quasi-likelihood estimates of the standard error.

A log-linear risk model was preferred over a logistic model, for the following reason. The 'risk' of being linked to a census record was comparatively high. Therefore, the odds ratio of linkage for a stratum with, say, 80% linked compared to a stratum with 60% linked ($[0.8/0.2]/[0.6/0.4] = 2.67$) is quite different from the risk ratio ($0.8/0.6 = 1.33$). It is the risk ratio that is of interest, as that will be the 'adjustment' required to the observed risk ratios (approximated by the odds ratios from logistic regression) in the cohort analysis.

4.1.2.1 Regression analyses for demographic strata combined

Regression analyses to determine the linkage bias due to small area deprivation (NZDep91) were conducted for all deaths combined, and separately by sex for 25–74 year olds for occupational class (NZSEI).

Model selection was conducted by using a combination of prior information, and a backward elimination strategy. Prior information was of two main types. First, the univariate and stratified analyses were used as a starting point to consider likely interaction and confounding. Second, each subsequent log-linear model built on previous models. For example, initial models were developed for the effect of sex, age, and ethnic group on the probability of being linked to a census record. This model was then used as the baseline to examine whether the socioeconomic variable of particular interest (eg, NZDep91 score) had any effect over and above the demographic covariates of sex, age, and ethnic group.

The general backward elimination strategy was as follows. As a first step, all main effects and first order interaction products were included in an initial model. The Wald Type III Chi-Square statistic and p value for each first order interaction product were then inspected. If the p value was statistically significant ($p < 0.05$) the interaction product was retained in the second step, otherwise the interaction product was discarded. (If three or more main effects were involved in two or more overlapping and statistically significant first order interaction products (eg, [age]×[sex], and [age]×[ethnic group]), then models with second order interaction products were explored.) In the second step, a model was fitted with the statistically significant first order interaction products from the first step, and all main effects. In the final step, the remaining statistically significant first order interaction terms and any main effects not involved in an interaction term were retained. Note that in the second and final step, main effects were retained even if not statistically significant – all main effects modelled (sex, age, ethnic group, time period between census and death, NZDep91 score, and NZSEI occupational class) were either the exposure of interest, or covariates that had strong prior justification for inclusion.

Often, the iterative estimation of the parameters of the log-linear model failed to converge when a number of main effects and their interaction products were included. In these instances, greater use was made of prior information, and exploratory logistic modelling to determine which interaction products to retain.

4.1.2.2 Regression analyses separately within the four sex by age groups

The previous regression strategy assumes that, unless rejected by statistical tests, the linkage bias by socioeconomic factors is homogenous across demographic strata. To complement this statistical approach, log-linear regression models were also conducted within each of the four sex by age groups used in the cohort analyses in this report (ie, 25–44 and 45–64 year old males, and 25–44 and 45–64 year old females). Further, these analyses excluded deaths in the first six months after census night, thus being more representative of the actual follow-up period for the cohort study. The models simply included dummy variables for five-year age groups and ethnicity (Maori, Pacific) as covariates, and a categorical variable for NZDep91 or NZSEI. No interaction terms were modelled as the analysis was already being conducted within groups of age by sex.

4.2 Cohort analyses

The cohort analyses refer to the association of socioeconomic exposures with the mortality outcome among the 25–64 year old 1991 census cohort. Analyses for 0–24 year olds, and 65–74 year olds are beyond the scope of this report.

The main epidemiological effect measure of interest was the *risk ratio*. The reason for using a *risk* rather than *rate* ratio was that the cohort was of short duration follow-up with a relatively rare outcome (death). Therefore an analysis using person–time in the denominator would give essentially the same result as that using counts in the denominator. (Estimating person–time in the NZCMS would also be done with some error – for example, it was not known which individuals completing a census record subsequently emigrated.) Moreover, it would have entailed a large (and probably not feasible) effort to have assembled the necessary data for a person–time analysis.

The cohort study-base was variously restricted for different analyses. Two main cohorts were used for the analyses: the *full cohort* and the *restricted cohort*. The full cohort consisted of all New Zealand resident and non-absent census records. (The SNZ census data set is hierarchically organised with individuals within dwellings within meshblocks. Some of the individual records are for people away from their usual residence on census night, with values for sex and age only. These records are called 'absentee records'. For each absentee record, one assumes that there is another fully completed census record elsewhere in the census data set within another dwelling for the same person. Thus, restricting the census data set to all non-absentee records should, theoretically, include one complete individual record for every person completing a census form.) The restricted cohort excluded those census records with missing data for any one of household income, car access, education, and labour force status. The major reason for a census record being excluded was missing data for household income. To have a valid household income required that all usual residents of the household aged 15 years and older be at home on census night, unless the total income of those adults actually at home on census night exceeded \$70,000.

4.2.1 Univariate analyses

The 'univariate' association of each socioeconomic exposure (small area deprivation, education, labour force status, occupational class, housing tenure, car access, and household income) with all-cause mortality was determined, separately for 25–44 and 45–64 year old males, and for 25–44 and 45–64 year old females. The main univariate results were presented for the restricted cohort. For each of the univariate analyses, the initial summary results are presented *excluding* all census records with a linked death in the first six months. That is, a priori, deaths in the first six months are assumed to be most at risk of health selection effects and were therefore excluded.

Results were presented *crude* using simple categorical analyses, and *adjusted* for age (five-year age groups) and ethnicity using logistic regression. Logistic regression analyses were conducted in SAS versions 6.12 and 8.0. All independent variables were specified as dummy categorical variables. As the outcome was relatively rare, the odds ratio closely approximated the risk ratio. Consequently, the terms risk ratio and odds ratio are used interchangeably in this report.

4.2.2 Sensitivity analyses of the univariate results

A range of sensitivity analyses were conducted for the univariate results to assess the likely impact of selection bias, health selection and misclassification of the mortality outcome. (The majority of these sensitivity analyses are reported in Appendix C of this report.) Age and sex adjusted logistic regression analyses were used for the majority of all-cause mortality sensitivity analyses. However, crude risk ratios were used for some of the sensitivity analyses (particularly cause-specific mortality analyses) to reduce the number of logistic regression models required. The use of crude data also allowed the use of 'floating' risk ratios (ie, whole population risk as reference risk). These floating risk ratios were sometimes easier for the interpretation of trends compared to logistic regression where a greater than expected change in the reference category risk may distort all other odds ratios. At several stages age and sex-adjusted logistic regression analyses were run as checks to ensure that trends emerging from the crude risk ratios and age and sex-adjusted odds ratios were consistent.

4.2.2.1 Selection bias

The census includes nearly all the New Zealand population. (The 1996 Post Enumeration Survey suggested a less than 1% undercount of the 1996 census among 25–64 year olds (Ewing 1997).) However, using the restricted cohort for the univariate (and latter multivariate) analyses may introduce some selection bias. A simple test of selection bias was, therefore, to repeat the univariate analyses where possible for the fullest cohort available for the given socioeconomic factor. For example, nearly the entire full cohort had a non-missing highest qualification.

4.2.2.2 Health selection

Many studies control for health selection effects by discarding the first few years of outcome data – a luxury not available in the NZCMS where deaths were only linked up to three years following the census. Therefore, considerable effort was expended in this report trying to determine how much bias may have arisen due to health selection in the NZCMS. The amount of health selection effect will vary for each socioeconomic measure (and for each health outcome measure). It should be most marked for labour force status, whereby people in poor health move into the non-active labour force. Household income is likely to be affected by health selection effects, but less so than individual income if the unhealthy person in the household is not the only or main income earner. For people older than 25, health selection effects should have no (or very little) effect on gradients in mortality by educational status, as education is by then a (nearly) fixed characteristic. Likewise, one would expect little health selection for gradients of mortality by small area deprivation in the short term.

Several strategies were available to examine the possible impact of health selection.

Strategy 1: Plot mortality risks over time by level of the given socioeconomic factor

If health selection effects exist, they would be expected to diminish over time (Fox and Goldblatt 1982; Fox et al 1985). Thus changes in the observed mortality risks were plotted over time for each level of a given socioeconomic factor. Depending on the socioeconomic factor, health selection would predict a convergence or divergence of the mortality risks over time. For example, if the income mortality gradient was biased by *drift* health selection, then one would expect to see a high mortality risk among low-income people initially, but it would fall over time as those with a low income consequent on their poor health either died or returned to good health. Conversely, the mortality risk among the high-income people would initially be very low as (under the health selection argument) one could only have a high income if in good health. But over time, the mortality risk among those with a high income would increase as some people succumbed to poor health. Thus, under the drift health selection hypothesis, one would expect to see a convergence of mortality risks by strata of income over time among the whole cohort (ie, a funnel plot). Alternatively, if the income mortality gradient was affected by *differential* health selection, one would expect a divergence over time in the mortality risks by income level among the active labour force. That is, as people from lower socioeconomic groups were supposedly more likely to be forced out of the labour force by poor health than people of higher socioeconomic groups, one would expect to see low mortality risk initially among low-income people in the labour force. But over time, the mortality risk for low-income people in the labour force would rise towards its 'background' risk. Conversely, the mortality risk for high-income people would not rise as much over time. Thus, the mortality risk plots by income-level for the labour force

would tend to diverge over time. Given that the NZCMS was a short-duration follow-up study, the plotting of mortality risks over time was only useful for investigating short-term health selection effects.

Strategy 2: Exclude recipients of sickness benefits

If *drift* health selection exists, one might expect the income mortality gradient among those people not receiving a sickness benefit (ie, excluding some of the ‘unhealthy’) to be reduced in magnitude. However, as sickness benefits are more likely to be received by poorer people for any given level of sickness (receipt of the benefit was means tested), restricting the cohort to non-recipients of sickness benefits may actually overadjust the income–mortality gradient for drift health selection. Assuming that both education and small area deprivation are unaffected by short-duration health selection drift, the baseline changes were established (ie, the ‘overadjustment changes’) to the education and deprivation mortality gradients after excluding sickness beneficiaries. For evidence that drift health selection was affecting the income mortality gradient, the reduction in the income gradient after excluding sickness beneficiaries had to be substantially larger than that observed for both education and deprivation.

Strategy 3: Exclude death outcomes that were hospitalised between 1988 and census night

It was possible to identify all deaths with a hospitalisation between 1988 and 1991 census night, and exclude them from the analyses. The rationale for this exclusion was that deaths among people with no hospitalisation might represent a subset of more ‘acute’ deaths, and thereby include deaths that were less prone to health selection effects on census night. There were four limitations to this approach. First, it would have been preferable to exclude all those census respondents who had been hospitalised in the three years prior to census night, thus restricting the cohort-base – not just restricting the outcomes. However, this would have required linking all hospitalisations for the 1991–94 period to the 1991 census – a massive task with small marginal gain to the NZCMS. Second, and like receipt of a sickness benefit, hospitalisation in the three years prior to census night was an imperfect marker of ill health. Not all people with poor health on census night would have been admitted to a (public) hospital in the preceding three years, and not everyone admitted would have been ill in the sense of being prone to health selection (eg, hospitalisation for injury). Third, and also similar to the receipt of a sickness benefit, hospitalisation for a given disease or health-state probably varies by socioeconomic position. For example, an unwell patient in a hospital’s emergency department may be more likely to be admitted if they do not have good social support at home, and social support may in turn be correlated with socioeconomic position (Berkman and Glass 2000). Moreover, private hospitalisations were not captured, which are also highly correlated with socioeconomic position. Fourth, the likelihood of hospitalisation prior to census night would vary by cause of death.

Nevertheless, excluding pre-hospitalised deaths was still a useful test of health selection. To maximise the validity of this sensitivity analyses, baseline analyses were again conducted for the socioeconomic factors where short-term health selection was assumed not to be important (small area deprivation, and highest qualification). As with receipt of a sickness benefit, for the exclusion of pre-hospitalised deaths for the income analyses to suggest health selection the change in the gradient had to be substantially more than that for the baseline deprivation and education analyses. Finally, greater attention was paid to cause-specific mortality.

4.2.2.3 Misclassification of the mortality outcome

The results from the linkage bias analyses were used to adjust the results from the cohort analyses as a sensitivity analysis for misclassification bias of the mortality outcome. For example, assume that decedents of lower socioeconomic position were 10% less likely to be linked to a census record than decedents of higher socioeconomic position. That is, the risk ratio from the linkage bias would have been 0.90 for the lower compared to the higher socioeconomic position groups. Further, assume that the risk ratio (approximated by the odds ratio) for the association of lower socioeconomic position compared to higher socioeconomic position was measured as 2.0 in the cohort analysis. The adjusted risk ratio would therefore be $2.0/0.9 = 2.22$. Note that the observed risk ratio was an underestimate by 10%, but that the *excess* risk ratio (ie, $RR - 1.0$) was underestimated by 18% ($0.22/1.22$).

4.2.3 Univariate analyses excluding the non-active labour force

Multivariate analyses that control for labour force status were problematic to interpret. Therefore, as a prelude to these multivariate analyses, univariate analyses were conducted that simply excluded the non-active labour force. The difference in effect size (1 minus the odds ratio) for the analyses excluding the non-active labour and the analyses including all labour force categories was then determined. A change in the effect size of a given socioeconomic factor with mortality after excluding the non-active labour force may represent one or more of six processes:

- variation (*effect modification*) of the association within the active labour force compared to all labour force categories, due to:
 - differential health selection
 - factors other than differential health selection
- controlling for *confounding* of the association by labour force status, where labour force status is a marker of:
 - health status (ie, drift health selection)
 - factors other than health status
- explaining that component of the association that is *mediated* by labour force status, where labour force status is a proxy for:
 - health status (ie, health status as an intermediary variable between socioeconomic position and mortality)
 - factors other than health status.

Thus, interpretation of the univariate results excluding the non-active labour force required cross-reference to results for sensitivity analyses of health selection, and results for other socioeconomic factors where one would expect a different balance of these processes to be working.

4.3 Multivariate analyses

Unless stated otherwise, all logistic regression results for the univariate analyses actually adjusted for age and ethnicity. Thus, the use of the term univariate was one of convenience. The term 'multivariate' in this report is reserved for those logistic regression analyses that include two or more socioeconomic factors (eg, education and income).

The multivariate models were specified and interpreted as best as possible with reference to the causal model shown in Figure 3, although problems with labour force status limited the full application of this simple framework. All independent variables were specified as dummy categorical variables. Categories were aggregated further compared to those for the univariate analyses to avoid problems with sparse data. The logistic regression models were conducted in SAS version 6.12 and 8.0.

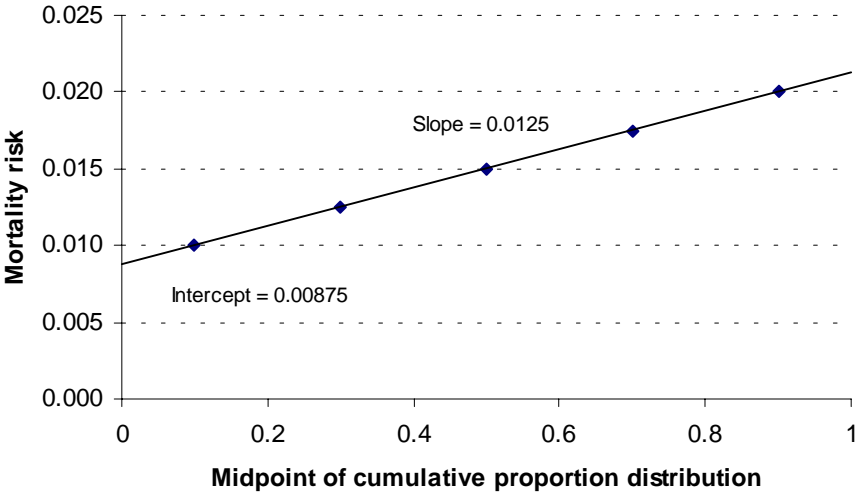
4.3.1 Relative index of inequality

To assist interpretation of the multivariate analyses, relative indices of inequality (RIIs) were calculated for income and education (Mackenbach and Kunst 1997). Conceptually, the RII is the relative risk of mortality for the person with the lowest socioeconomic position compared to the person with the highest socioeconomic position. As such, it assumes that socioeconomic factors simply *rank* individuals within a society. For example, assume that income was measured as a quintile variable, and that the mortality risk in each quintile (from highest to lowest income) was 0.010, 0.0125, 0.015, 0.0175, and 0.020. That is, the relative risk for people living in the poorest income quintile (0.020) compared to those in the richest income quintile (0.010) was 2.0. This relative risk compares two groups, and is thus essentially a comparison of the average mortality risk about the 10th percentile compared to the 90th percentile of the population. For two reasons, it may be useful to consider the mortality risk for the poorest (zero percentile) compared to the richest (100th percentile):

- intrinsically, it is of interest to estimate the gradient right across the socioeconomic hierarchy, rather than just comparing groups
- often the groupings vary in size between studies. For example, in one study half the population may be in the reference category of education, whereas in another study only 10% of the population may be in the reference category. Thus, comparing the effect measures between these two studies is confounded by differences in group sizes.

Figure 8 below demonstrates how the RII is calculated, using the hypothetical mortality risks given above. The income quintiles are ranked from the highest socioeconomic group to the lowest. Each quintile comprises 20% of the population. Thus, the richest quintile is plotted at 0.1 on the cumulative proportionate distribution of the population (x axis), with a mortality risk of 0.01. The next richest income quintile will have an x-axis value of 0.3 (0.2 for the previous quintile, plus half of the current quintile), and a y-axis value of 0.0125, and so on.

Figure 8: Hypothetical example of mortality risk by income to demonstrate the calculation of the RII



Having plotted these x-y points, the slope and intercept can be calculated. In this simple example, the slope is 0.0125 and the intercept is 0.00875. The RII is then $(0.0125 + 0.00875)/0.00875 = 2.43$. That is, the poorest person has a mortality risk that is 2.43 times that of the richest person, somewhat more than the relative risk of 2.0 derived from simply comparing the lowest and highest income quintiles.

In this report, RIIs are calculated for age and sex adjusted analyses, and for multivariate analyses. Therefore, odds ratios were used as the y-variable. (Given that mortality was a relatively rare outcome, the odds ratios are directly proportional to the adjusted mortality risks.) Weighted linear regression was conducted of the odds ratios on the midpoints for each income and education group on the cumulative proportion distribution. Weights were the inverse of the variance of the crude mortality risk for each income or education group (ie, $p(1-p)/n$ (Kirkwood 1988), where n was the census count and p the mortality risk/proportion). (A more statistically 'correct' way to calculate the RIIs would have been to rerun all the logistic regressions specifying income and education as continuous variables ranging from zero to one.)