Public Health Monograph Series

No. 4

ISSN 1173-6844

Anonymous record linkage of 1991 census records and 1991-94 mortality records

The New Zealand Census-Mortality Study

(NZCMS Technical Report No. 1)

Tony Blakely

Clare Salmond

Alistair Woodward

December 1999

Department of Public Health,

Wellington School of Medicine

ISBN 0-473-06700-5

Published by the Department of Public Health
Wellington School of Medicine
PO Box 7343
Wellington South
Wellington
New Zealand

ISBN 0-473-07751-5

Acknowledgements

Staff of Statistics New Zealand, in particular:

- Paul Willoughby and Keith McLeod who undertook the actual record linkage, and monitored privacy issues on a day-by-day basis
- Sandra McDonald who managed the Data Laboratory access
- Sharleen Forbes, John Cornish, and Tracey Gilmour who all provided managerial input
- Robert Didham who answered detailed questions regarding the 1991 census data set
- Len Cook, Government Statistician, who provided the initial approval for the New Zealand Census-Mortality Study to proceed.

Co-investigators in the New Zealand Census-Cohort Mortality Study, of which this record linkage project is the first step: Peter Davis, Cindy Kiro, and Neil Pearce.

Colleagues from the Department of Public Health, Wellington School of Medicine, in particular Philippa Howden-Chapman and Peter Crampton.

Staff of New Zealand Health Information Services, in particular Tracey Stewart, Jim Fraser, Liz Mooney, and Barbara Bridger.

Table of Contents

Tables

Figures

Glossary

Absolute odds — The absolute odds is derived from the relative odds by allowing for the expected number of true links and the size of the two files. It is similar to a betting odds for any one comparison pair being a true link. The absolute odds logarithm to base two (the total absolute weight) is more simply derived than the absolute odds - see 'total absolute weight'.

Absolute weight method — A method to estimate the number of false positives at a given probabilistic weight in the record linkage. Whilst precise, it is prone to bias from correlated agreements and disagreements among matching variables.

Agreement frequency ratio — The odds of the [probability of agreement on a matching variable for true links] to the [probability of agreement on a matching variable for true non-links] - ie. the ratio of the *m* probability to the *u* probability. It may be a *global agreement frequency ratio* (eg for all values of year of birth) or a value *specific agreement frequency ratio* (eg for each value of year of birth).

Agreement weight — The logarithm to base two of the agreement frequency ratio - ln[m/u] / ln[2].

Best link record linkage — Record linkage where each record can only be linked once. For example, in this research each mortality record could only be linked with one census record.

Blocking — A procedure used in record linkage to reduce the number of possible comparisons. That is the records on both files are divided into blocks (eg area of residence), and record linkage is conducted within these blocks only.

CP pair — Clerical review pair - a pairing of a mortality and a census record that is between the clerical review cut-off and the match cut-off, i.e. a pair that must be viewed to make a decision as to whether it is to be accepted or rejected.

Duplicate method — An method to estimate the number of false positives above a given weight range for record linkage where one to one linking only is allowed.

DA pair — Duplicate pair in 'A' file (Automatch® label for census file) – a pairing of a census and a mortality records where the *mortality* record is also involved in another MP or DA pair at or above the DA pair's weight. That is, the mortality record is paired with more than one census record at or above the DA pair's weight.

Data-set (Database)	"An organised set of data or collection of files that can be used for a specified purpose."[1] A database may consist of one file, or more relational files. In this research, the linkage of mortality and census files may be thought of as creating a linked data-set.
DB pair	Duplicate pair in 'B' file - a pairing of a census and a mortality records where the *census* record is also involved in another MP or DB pair at or above the DB pair's weight. That is, the census record is paired with more than one mortality record at or above the DB pair's weight. DB pairs were uncommon in this research because the census file was so much larger than the mortality file.
Disagreement frequency ratio	The odds of the [probability of disagreement on a matching variable for true links] to the [probability of disagreement on a matching variable for true non-links] - ie. the ratio of 1 minus the *m* probability to 1 minus the *u* probability. It may be a *global disagreement frequency ratio* (eg for all values of year of birth) or a value *specific disagreement frequency ratio* (eg for each value of year of birth).
Disagreement weight	The logarithm to base two of the disagreement frequency ratio, ln[(1-m)/(1-u)] / ln[2].
Chance method	A method to estimate the number of false positive links above a match cut-off when most links above the match cut-off agree exactly on all matching variables. It is based on the estimated average probability of any one mortality record and any one census record agreeing exactly (for matching variables only), purely by chance.
False positive links (false links)	The estimated number of accepted links that are not true links.
Field	The information as presented in a file for each variable. For example, the income field in the census file is the information for the variable income for each record (or person). Fields are often represented by columns in a computerised file.
File	A collection of variable information for multiple units of observation (each unit of observation comprising one record).
Frequency ratio	The *agreement* frequency ratio is the odds of a particular matching variable agreeing in true links versus true non-links, that is the *m* probability divided by the *u* probability. The *disagreement* frequency ratio is [1 minus the *m* probability] divided by [1 minus the *u* probaility]. The frequency ratio can be calculated globally for all values of the matching variable (global frequency ratio), or separately for each possible value of the matching variable (specific frequency ratio). The latter increases the discriminatory power for uncommon values of the matching variable.

Agreement frequency ratio	(See 'frequency ratio'.)
Disagreement frequency ratio	(See 'frequency ratio'.)
Log-linear risk model	Regression model with a logarithmic link. In this project, the dependent variable was binary (0 or 1) resulting in model predicting the proportion of mortality records linked for each combination of independent variables (eg age, occupational class). A binomial error structure was specified.
Label	The name for a field in a file. For example, total individual income (or more precisely TINC91) is the label for the field containing total income on each individual in the census.
Link	A mortality and census record pair that is accepted as likely to be a true link, either by being above the match cut-off or by being accepted on clerical review.
***m* probability**	The probability that a variable agrees given that the comparison pair being examined is a true non-linked pair. It may be global (eg for all values of year of birth) or value specific (eg for each value of year of birth). It is initially specified by the user of Automatch®, but can be better approximated on a preliminary set of links by calculating the proportion of B file records that agree with a particular value for a particular variable of the A file records.
Match specification	The specification of the blocking variable, the matching variables (eg. *m* and *u* probabilities, partial agreements, arrays), the match and clerical review cut-offs for a single pass.
Matching variable	The variables that are common to the two files being linked, and are used in the probabilistic linkage of records.
Match-run	The full sequence of passes used in the record linkage.
Meshblock	The smallest geographic area used for coding purposes by Statistics New Zealand, with a median size of 90-100. The meshblock code assigned to most census and mortality records in this study was the most discriminating variable in the record linkage.
MP pair	Match pair - a pairing of a mortality and a census record that is above the match cut-off in Automatch®, i.e. a link.
Pass	The comparison of records on two files within a single match specification.
Positive predictive value	The percentage (or proportion) of accepted links that are true links.

Record	All the variable information for one unit of observation in a file. For example, one record from the census file would include all the variable values for an anonymous person (names and text address are not available). Records are often represented by rows in a computerised file.
Relative odds	The product of the agreement and disagreement frequency ratios for all matching variables for any comparison pair of records.
Sensitivity	The proportion (or percentage) of all true-links (detected and undetected) that have been detected as links.
Total (combined) relative odds	(See 'relative odds'.)
Total (combined) relative weight	The sum of the agreement and disagreement weights for all matching variables for any comparison pair of records. Equivalently, it is the logarithm to base two of the total (combined) relative odds.
Total absolute odds	(See absolute odds.)
Total absolute weight	The logarithm to base two of the absolute odds (the betting odds of a comparison pair being a true link). The absolute odds is derived from the relative odds allowing for the expected number of true links and the size of the two files.
True links	The (theoretical) set of links where the census and mortality record are for the same individual.
True non-links	The (theoretical) set of links where the census and mortality record are *not* for the same individual.
***u* probability**	The probability agreement on a given matching variable among true non-links (i.e. that probability that variables agree purely by chance between non-links). It may be global (eg for all values of year of birth) or value specific (eg for each value of year of birth). It is approximated in Automatch® by the proportion of all records in both files that have a particular value for a particular variable (i.e. value specific).
Weight	In record linkage using Automatch®, the weight is the logarithm (base two) of the frequency ratio.

Abbreviations

CAU	Census Area Unit (median population about 2,000)
dd	day of birth
E[FP]	Expected number of false positive links
mm	month of birth
NHI	NZHIS's National Health Index - files personal identifier data for nearly every individual in New Zealand, and links separate NMDS events for the same individual by means of a unique identifier (the NHI number)
NMDS	NZHIS's National Minimum Data-Set - files data on both hospitalisation events and death events
NZHIS	New Zealand Health Information Services
NZSCO-68	New Zealand Standard Classification of Occupations, 1968 version
NZSCO-90	New Zealand Standard Classification of Occupations, 1990 version
NZSEI	New Zealand Socio-Economic Index (an occupational class index)
OPCS	Office of Population Censuses and Surveys (United Kingdom)
PPV	Positive predictive value
SNZ	Statistics New Zealand
yyyy	year of birth

Chapter 1: Introduction

1.1 Why link census and mortality records?

This technical report describes the anonymous record linkage of 1991 New Zealand census records with 1991-1994 mortality records. The reason for this linkage was to create a cohort study of the 0-74 year old New Zealand population followed up for mortality for three years. That is, we were seeking to take the entire 0-74 year old New Zealand population on census night 1991, and determine who died in the following three years. Why create this cohort study? Because the census has good measures of socio-economic status (e.g. income, education, occupation, car access and housing tenure) that allows a comprehensive study of the association of socio-economic factors with mortality in New Zealand. Pearce and colleagues have determined the association of occupational class with mortality for 15-64 year olds in New Zealand for 1974-78, 1985-87, and (unpublished as yet) 1995-97.[2-11] There are several advantages from creating a census cohort study to measure socio-economic mortality gradients compared to the occupational class analyses by Pearce and colleagues:

- The analyses by Pearce and colleagues used death data for the numerator occupational class, and census data for the denominator occupational class. Mortality and census data elicit occupation with different questions (usual versus current, respectively) and from different people (next of kin versus self-identified, respectively). Whilst the 'unlinked' analyses by Pearce and colleagues are likely to be a fairly accurate representation, a linked cohort study allows an alternative analysis.
- The range of subjects can be extended beyond males aged 15-64 to include both sexes and all age groups. (We have restricted the record linkage to 0-74 year olds for two reasons: at ages greater than 74, there is likely to be increased residential mobility (e.g. moving to resthomes) limiting our ability to link mortality and census records; it is generally accepted that socio-economic mortality gradients are stronger and more policy relevant for the young and middle aged.)

- The range of exposures can be extended beyond occupational class to include income, education, car access, and housing tenure. Furthermore, household socio-economic characteristics can be examined (e.g. household income, and highest, average, or head of household occupational class).
- There is enough power to consider multi-level causation (e.g. the effect of small area deprivation, over-and-above personal socio-economic status, on mortality risk).
- As the census and mortality data are combined for each death, it is possible to determine the numerator-denominator bias in ethnic group recording on mortality compared to census data – a problem that has plagued the comparison of ethnic specific mortality rates in New Zealand.[1]

In addition to existing work on occupational class and mortality, there is a growing body of work in New Zealand looking at socio-economic mortality gradients by small area (neighbourhood) deprivation.[12-14]. However, small area deprivation is not strictly a personal or individual-level measure of socio-economic status. A census cohort study allows the examination of a range of individual-level socio-economic factors, *and* a determination of the independent effects of small area deprivation over-and-above individual-level socio-economic factors.

1.2 International precedents

Census and mortality records have been linked in at least four other countries.

A one percent sample of the 1971 census in England and Wales (n=500,000) formed the basis of the OPCS Longitudinal Survey.[15] These census records were then linked to subsequent health, mortality and immigration data. In excess of 95% of the deaths between 1971-75 for this cohort were estimated to have been successfully linked.[15] Data available for record linkage in the OPCS included names.

[1] Routinely published mortality rates for Maori and nonMaori are calculated by using the health data to determine ethnic specific numerators, and census data to determine ethnic specific denominators. It is well established that health and census data collect ethnic group differently – thus there is bias in the routinely published ethnic group mortality rates.

In Italy, 78.2% of 120,436 deaths in the six month period following the 1981 census were linked to a census record.[16] Exact methods were not stated by the authors, but the linkage did use anonymous census records and “census identification codes” - presumably, a probabilistic record linkage method was used.

Whilst not using a one-off census of the entire population, the National Longitudinal Mortality Study (NLMS) in the United States followed 1 million respondents of the Current Population Surveys.[17] This cohort was linked to the National Death Index using probabilistic record linkage methods and clerical review of ‘questionable’ links – name was available on both files. In a validation study using a similar probabilistic record linkage methodology to the NLMS, all deaths were detected and only 1% of accepted links were incorrect (i.e. the positive predictive value of the record linkage was 99%).[18]

The Swedish have a strong history of linking census and mortality data to facilitate research on socio-economic inequalities in mortality. For example, deaths for 1961 to 1979 linked to the 1960 census were used to analyse death by social class,[19] and likewise deaths for 1986 to 1990 linked to the 1985 census have been used to analyse avoidable mortality.[20] These linkages are possible because of the widespread use personal identifier codes since the 1950s. In an elegant study that demonstrates the power of record linkage, Leon et al (1998) link 15,000 birth records from 1915-29 to their subsequent 1960 and 1970 Swedish census records and death records (if applicable).[21] This study provided persuasive evidence that fetal growth rate is etiologically important for later ischaemic heart disease.

1.3 Privacy, and SNZ Security Statement

Never before has the New Zealand census been linked to an external data set. An over-riding concern in any such linkage is the protection of privacy. The New Zealand census is administered by Statistics New Zealand (SNZ), who are required to function within the bounds of the Statistics Act (1975) regarding access to unit record census (and other) data by researchers. There are also wider concerns regarding census data.

The New Zealand census has a high completion rate, and is an essential information source for planning in New Zealand. To obtain accurate data from as many people as possible, it is important that individual privacy is protected for the census, so that the confidence of the New Zealand populace in the census can be maintained at current high levels. To maintain individual privacy of census data, and meet the requirements of the Statistics Act (1975), a comprehensive process was put in place for this project – this is outlined in a security statement issued by Statistics New Zealand and presented in the box below.

STATISTICS NEW ZEALAND SECURITY STATEMENT

The New Zealand Census-Mortality Study was initiated by Dr Tony Blakely and his co-researchers from the Wellington School of Medicine, University of Otago. It was approved by the Government Statistician as a Data Laboratory project under the Microdata Access Protocols.

Requirements of the Statistics Act

Under the Statistics Act 1975 the Government Statistician has legal authority to collect and hold information about people, households and businesses, as well as the responsibility of protecting individual information and limits to the use to which such information can be put. The obligations of the Statistics Act 1975 on data collected under the Act are summarised below.

1. Information collected under the Statistics Act 1975 can be used only for statistical purposes.

2. No information contained in any individual schedule is to be separately published or disclosed to any person who is not an employee of Statistics New Zealand, except as permitted by sections 21(3B), 37A, 37B and 37C of the Act.

3. This project was carried out under section 21(3B). Under Section 21(3B) the Government Statistician requires an independent contractor under contract to Statistics New Zealand, and any employee of the contractor, to make a statutory declaration of secrecy similar to that required of Statistics New Zealand employees where they will have access to information collected under the Act. For the purposes of implementing the confidentiality provisions of the Act, such contractors are deemed to be employees of Statistics New Zealand.

4. Statistical information published by Statistics New Zealand, and its contracted researchers, shall be arranged in such a manner as to prevent any individual information from being identifiable by any person (other than the person who supplied the information), unless the person owning the information has consented to the publication in such manner, or the publication of information in that manner could not reasonably have been foreseen.

5. The Government Statistician is to make office rules to prevent the unauthorised disclosure of individual information in published statistics.

6. Information provided under the Act is privileged. Except for a prosecution under the Act, no information that is provided under the Act can be disclosed or used in any proceedings. Furthermore no person who has completed a statutory declaration of secrecy under section 21 can be compelled in any proceedings to give oral testimony regarding individual information or produce a document with respect to any information obtained in the course of administering the Act, except as provided for in the Act.

Census data

The Population Census is the most important stocktake of the population that is carried out. The statistics that are produced provide a regular picture of society. Results are used widely in making decisions affecting every neighbourhood. They are used in planning essential local services, and they also help to monitor social programmes ranging from housing to health.

Traditionally census data is published by Statistics New Zealand in aggregated tables and graphs for use throughout schools, business and homes. Recently Statistics New Zealand has sought to increase the benefits that can obtained from its data by providing access to approved researchers to carry out research projects. Microdata access is provided, at the discretion of the Government Statistician, to allow authoritative statistical research of benefit to the public of New Zealand.

This project used anonymous census data and mortality data which were integrated using a probabilistic linking methodology to create a single dataset that allows the researchers to undertake a statistical study of the association of mortality and socio-economic factors. This is the first time that the census has been linked to an administrative dataset for purposes apart from improving the quality of Statistics New Zealand surveys. The project has been closely monitored to ensure it complies with Statistics New Zealand's strict confidentiality requirements.

Further information

For further information about confidentiality matters in regard to this study please contact either:

Chief Analyst, Analytical Support Division, or
Project Manager, Data Laboratory

Statistics New Zealand
PO Box 2922
Wellington

Telephone: +64-4-495 4600
Facsimile: +64-4-495 4610

1.4 Objectives

The record linkage reported in this technical report is the first step of a larger project – the New Zealand Census-Mortality Study (NZCMS). The NZCMS aims to determine the association of socio-economic factors measured on the census with mortality across four census-cohorts – 1981, 1986, 1991, and 1996 censuses. The record linkage of the 1991 census with 1991-94 mortality records described in this technical report was a pilot study with the **aim**:

- to determine the feasibility of anonymously linking census and mortality records using probabilistic record linkage software (Automatch®).

Anonymous record linkage was necessary as the New Zealand census data base does not include text names and address – the only personal identifiers that could be used as matching variables were day, month, and year of birth, sex, ethnic group, country of birth, and (most importantly) a geocode for usual residence.

Beyond this general feasibility aim of the pilot, there were specific **objectives**:

- to determine the percentage of mortality records that could be linked to a census record.
- to determine the accuracy of the linkage as measured by the positive predictive value (i.e. the percentage of all links accepted that were estimated to be correct links of the same individual's census and mortality record).
- to analyse the variation by demographic and socio-economic factors, between linked and unlinked mortality records, and hence estimate the bias in the record linkage (i.e. the relative difference in probability of linkage for a high compared to low socio-economic individual).

A **target** was set by the investigators and staff of SNZ for the record linkage, namely that:

- at least 70 percent of mortality records for deaths six to 18 months following the census should be successfully linked to a census record

or

- 60 to 70 percent of deaths six to 18 months following the census should be successfully linked to a census record, with little apparent bias by socio-economic factors between linked and unlinked mortality records.

The structure of this technical report is largely built around the three specific objectives. Chapter 3 (Results – record linkage) addresses the first two objectives, and Chapter 4 (Results – analysis of bias) addresses the last objective.

1.5 Intended audience of this technical report

We envisage that this technical report will have a fairly specific audience. First, we (co-investigators of the NZCMS and staff of SNZ) will use this report as a guide to future linkage of other censuses with mortality data. Second, the technical report should be of use to people looking to link other routine data sets in New Zealand, and internationally. Third, people who read future output of the NZCMS may wish to know more about the linkage methodology and output.

As a summary of the detailed material in this report, we encourage readers to also refer to a forthcoming publication in the Australian and New Zealand Journal of Public Health (Blakely T, Woodward A, Salmond C, Statistics New Zealand. In press. Anonymous Linkage of New Zealand Mortality and Census Data.)

Chapter 2: Methods

2.1 Principles of probabilistic record linkage

Newcombe (1988) and Baldwin et al (1987) provide good basic introductions to probabilistic record linkage methods.[22] The method has been used in the United States to link mortality records to the Current Populations Survey database.[17] In the last decade, Jaro (1995) has developed an advanced software package, Automatch®, for probabilistic record linkage.[23] 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 power. Probabilistic record linkage formalises these intuitive process, using probability ratios and taking advantage of the processing capacity of computers.

2.1.1 Frequency ratios, *m* and *u* probabilities, and weights

At the heart of probabilistic record linkage are frequency ratios, determined by the *m* and *u* probabilities. Consider the variable day of birth (dd), and the value 9. The *m* probability is the probability among the **true 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 *u* probability is similar to the *m* probability, except it applies to the **true non-linked** records.

Without knowing which are the actual true linked and true non-linked records, the user must initially specify the *m* probabilities on the basis of a best guess of the likely

agreement for a given variable among 'true links'. After a preliminary pass (a comparison of records from both files using one strategy) or a preliminary match-run (one or more passes), value specific *m* probabilities for each variable can be estimated by simply calculating the observed proportion of accepted links that agree on a particular value for a particular matching variable. This is conducted in Automatch® by the MPROB function. For example, the *m* probability for ethnic group may be guessed at 0.80 for an initial pass - that is 80% of true links are guessed to agree on the ethnic group variable *regardless of the value of the ethnic group variable*. After the initial pass, it might be found that for links accepted thus far with an ethnic group of 'European' for the census record, the mortality record is also 'European' in 95% of links. However, it may be that the equivalent percentage for 'Maori' is 60% - a not unrealistic expectation given the poor coding of Maori ethnic group on health data compared to census data in New Zealand. Such a variation in the *m* probabilities by value is important to harness as it increases the discriminatory power of the record linkage.

Just as with the *m* probabilities we cannot measure the *u* probabilities directly as we do not know the true non-links. However, they can be closely approximated by the frequency of each specific value of each matching variable in the two files. In Automatch® this is calculated by simply pooling all the mortality and census records, and calculating the proportion of all pooled records that have each specific value of each variable. For example, each value of dd will have a frequency of about 0.033 (one divided by 30). As such, the *u* probability measures how likely any mortality record is to agree with any census record on the value of a variable *purely by chance*.

The example in Table 1 gives frequency ratios for agreement and disagreement on dd. The agreement frequency ratio of 32 to 1 for an observed match on dd is the odds of [the probability of dd matching in **true links**] to [the probability of dd matching in **true non-links**]. That is, dd is 32 times more likely to agree among true links than among true non-links. Conversely, the disagreement odds of dd not matching in true links versus true non-links is 1 to 19.

Table 1: Example of frequency ratios and weights for matching by the variable day of birth

Comparison outcome	Proportion/ Frequency True links	True non-links	Frequency ratio	Weight
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.

The frequency ratios presented in Table 1 are what Newcombe defines as the *global frequency ratios*. That is, the average frequency ratio for agreement, or disagreement, of all values for any particular variable. But the global frequency ratio does not allow for how common values of a given variable are. For example, whilst the frequency distribution of dd is even (meaning that the specific frequency ratios for each value of dd will differ little from the global frequency ratio), the frequency distribution of name is not. A match between two records on a rare name is more discriminating than a match between two records for a common name. The former (rare name) equates to a relative fall in the *u* probability, and hence an increase in the *specific frequency ratio* for agreement. A common name equates to a relative increase in the *u* probability, and hence a decrease in the *specific frequency ratio* for agreement. The use of specific frequency ratios increases the discriminatory power in probabilistic record linkage, more so if both value specific *m* and *u* probabilities are used.

Having determined the frequency ratios for each matching variable (and each value of each matching variable), the next step is to calculate the combined relative odds for any given comparison pair of, say, one mortality record and one census record. The combined relative odds 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 relative odds quickly becomes very large (for agreement) or very small (for disagreement), and it is easier to use the sum of the *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 1. 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.[24] A convenience of using logarithms to base two is that each increase in the weight by one represents a doubling of the overall odds. The combined weight can be used to allocate, by means of a cut-off, each possible comparison pair to one of a set of highly likely links, a set of highly unlikely links, and perhaps a set of uncertain links for clerical review. A comparison pair with a high combined weight is likely to be a true link; a comparison pair with a low weight (usually negative) is unlikely to be a true link.

2.1.2 Blocking

For a full comparison of two files each with 1000 records, there are 1000×1000 = 1,000,000 possible pair comparisons. Yet there may only be 1000 true links if the links are only one to one, and each record has a true link in the other file. Comparing all records on one file with all the records on the other file is computationally inefficient, as the vast majority of comparisons will be non-links. Therefore, *blocking* is used.

Blocking is a key step in probabilistic record linkage. It 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 100×10×10 = 10,000 with blocking by age ([blocks]×[records in each block in first file]×[records in each block in second file]).

If a variable is incorrectly recorded in one file, and that variable is used as the blocking variable for a particular pass, then it is impossible for that record to be correctly linked on that pass. This is known as *skipping*. For example, blocking by meshblock of usual residence will result in skipping of links due to either: a) inaccurate recording of meshblock on one (or both) files; or b) a change of residence for the individual concerned between the determination of usual residence of one file and that on the other file. To overcome skipping, more than one pass should be used, and each pass

should use a different (and preferably independent) blocking variable. A full sequence of passes using different blocking strategies is called a match-run. Extending the above example, a subsequent blocking strategy using components of date of birth and occupation would partition the files in a manner such that skipping on the first pass was independent of that on the second pass. (Unfortunately, in this research there was not enough discriminatory power for efficient blocking strategies other than variations on usual residence - the meshblock and larger census area unit. For example, blocking the files by day of birth would only create 31 blocks in each file, meaning numerous comparisons in each block, with both increased computational requirements and the increased probability of false positive links.) Efficient record linkage results from using the blocking variables that partition the files into as many blocks as possible in the first pass, and less restrictive blocking variables in later passes.

Further discriminatory power can be obtained by the use of partial agreement weights. For example, it is common for year of birth, and hence the derived age, to be reported and entered incorrectly by one or two years. An absolute difference in year of birth between two records of only one or two years is not as bad as a difference of, say, 20, 50 or more years. Therefore, it may be appropriate to assign a partial agreement weight for 'minor' disagreements. However, the calculation of these partial agreement frequency ratios, and hence the weight, is computationally difficult if done using odds. Automatch® does not assign partial agreement weights per se, but instead uses approximations to them due to the computational complexity. For example, one of the matching commands in Automatch® is PRORATED. This allows an absolute variation in numeric matching variables by a specified amount. The PRORATED command may be used for year of birth, with a specified tolerance of two years. The weight assigned for an exact match on age will still be the full agreement weight. But a mismatch by one year will be assigned a weight of:

$$[\text{full agreement weight}] - \frac{[\text{full agreement weight}] - [\text{full disagreement weight}]}{3}$$

That is, it will place the weight for a mismatch of one year one third of the way towards the full disagreement weight. Similarly, a mismatch of two years will receive a weight two thirds of the way towards the full disagreement weight.

2.2 Data used in the record linkage and analysis of bias

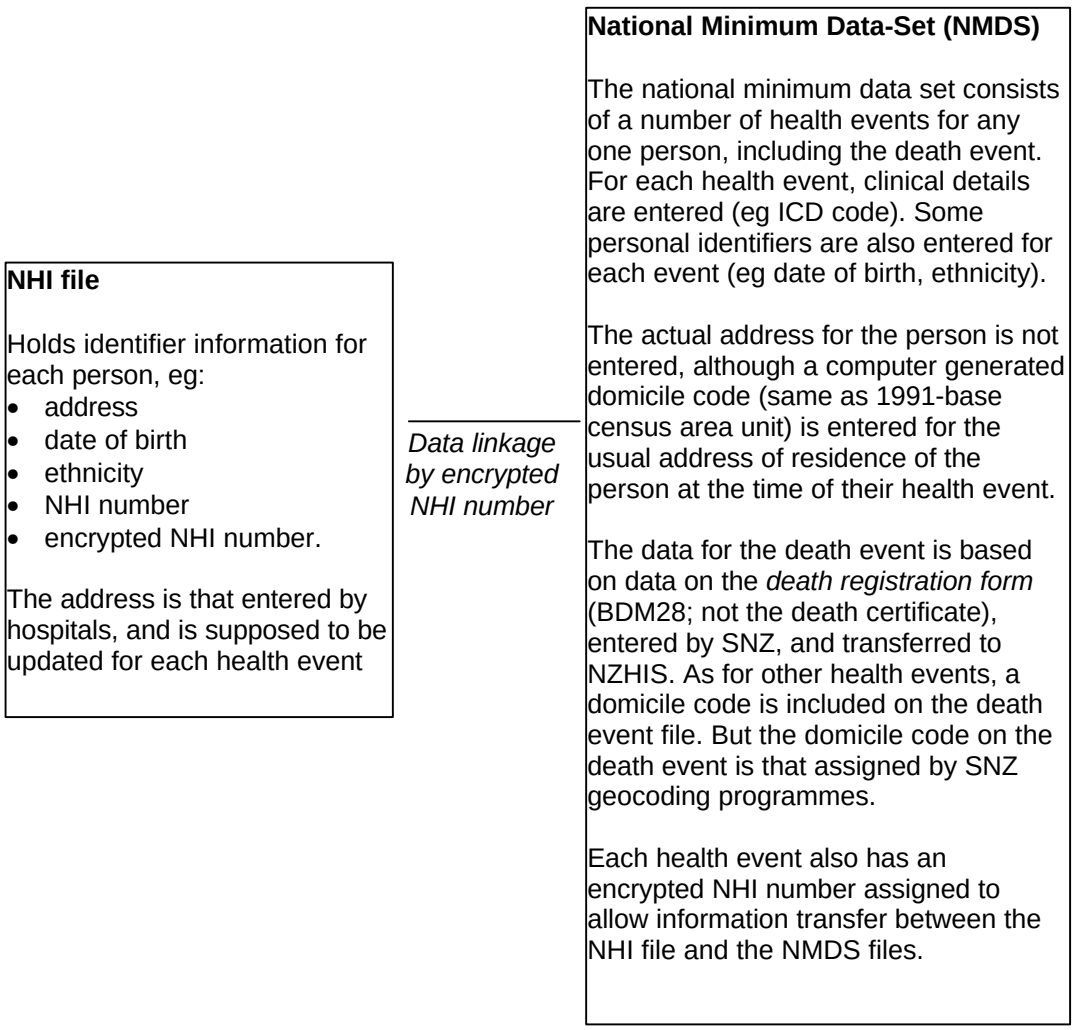
2.2.1 Mortality data

2.2.1.1 Structure of NZHIS data

The structure of New Zealand Health Information Services (NZHIS) data sets relevant to this research are presented in Figure 1. The important points to note are that:

- the mortality data for this research was stored in two data-sets - the National Health Index (NHI) file, and the National Minimum Data-Set (NMDS)
- the encrypted NHI number is included on both the NHI file and NMDS, allowing linkage of these data-sets
- the address (text) is held on the NHI file only, and should be the address for the last health event entered directly by a hospital. Therefore, the address should be that on the death registration form when the death occurred in hospital, but may not be for a death that occurred outside a hospital
- for any death event where there was no NHI file for that individual, the information on the death registration form was used to construct the NHI file
- separate domicile codes (NZHIS equivalent of SNZ census area code) are available for any one individual from their NHI record, and *each* of their NMDS health event records
- some variables are entered both on the NHI file and on the NMDS health event files (eg date of birth, ethnic group), and may differ between files for the same individual (eg changing self-defined ethnic group over time, coding errors for date of birth). Importantly, *the demographic data on the NHI file and the NMDS death event are independent for the majority of decedents*: the NHI File is usually data collected by a hospital clerk; the NMDS Death Event File is data collected by an undertaker on the BDM28 death registration form.

Figure 1: How health data is stored for anyone person by NZHIS



2.2.1.2 Mortality records obtained from NZHIS

The reason for doing the record linkage was to find the death events ('outcomes') for the 1991 census file, and so allow a short duration cohort study of all New Zealanders that completed a census questionnaire in 1991. The 1991 census provided measures of the socio-economic 'exposures' for the cohort study. The selected cohort was all individuals having completed the 1991 census, and aged 0-74 years on census night (5 March 1991). Therefore mortality records were requested for people dying between 5 March 1991 and 5 March 1994 inclusive, *and* aged 0-74 years of age inclusive on 5 March 1991. Mortality records with a NMDS Death Event File domicile code of 9999

(overseas resident) were discarded. A total of 42,229 mortality records were obtained from NZHIS.

For privacy reasons, and conditions of data access to the census under the Statistics Act 1975 (see SNZ Security Statement, page 19), it was inappropriate to transfer the NHI number, even in encrypted form, to SNZ with the mortality records. (Although the encrypted NHI number aims to protect individual confidentiality, it would still be possible that access to the final linked mortality and census record files, by someone with another file with both encrypted NHI numbers and actual names on it, may lead to disclosure of individual identity.) Therefore, a master mortality file was archived by NZHIS (CD Rom) with variables in addition to those transferred to SNZ (eg encrypted NHI number, actual address) for possible future reference. The data transferred from NZHIS to SNZ did not contain these identifying variables.

2.2.1.3 Mortality variables used in the record linkage and subsequent analysis of bias

The mortality variables used in the record linkage, and the subsequent analysis of bias in the record linkage, are presented in Table 2. There were four variables for which information was extracted from both the NHI file and NMDS death event file: date of birth (DOB), sex, ethnic group, and domicile code.

NZHIS did not store meshblock codes on either the NMDS or NHI files, just domicile codes (the equivalent of SNZ CAU codes). An important discovery early in this research was that the SNZ Vitals file did store the meshblock for about 90% of deaths from 1980 - the other 10% were assigned a CAU only as the exact meshblock was not determined. (The SNZ Vitals file is SNZ's own file of deaths and births. For deaths, SNZ enters the data directly from the death registration form and then forwards this database to NZHIS, thus forming the base for all NZHIS mortality files.) Three mortality variables - registration year, registration office, and registration number - when concatenated, made a unique identifier for all mortality records. All that was required to obtain meshblocks for the usual residence of the decedent at death was to create this concatenated variable, and merge the NZHIS mortality records with the SNZ Vitals file.

Three domicile (or CAU) codes were available directly from NZHIS for the project mortality data. First, that recorded on the NHI file (NHI-CAU). Second, it was possible to obtain a domicile code for some decedents for a health event before census night (pre-CAU). If the decedent had more than one health event before census night, then the domicile code for the health event immediately preceding the census was extracted. Third, it was possible to obtain a domicile code for some decedents for a health event after census night (post-CAU). As for pre-CAU, if there was more than one health event after census night, and before death, then the domicile code for the health event immediately after census night was used. In addition, there was the CAU created by aggregating the meshblock obtained from the SNZ Vitals file (Vitals-CAU). Together, these four CAU codes gave multiple options for blocking variables in the record linkage. In particular, note that the intention of the pre-CAU and post-CAU codes was to bracket as narrowly as possible the decedent's usual residence on census night, thus facilitating matching with the usual residence geocode of the same individual on census data.

Table 2: Mortality variables used in the record linkage and analysis of bias

Variable	Purpose	Comments
Date of Birth (NHI)	• matching variable	Date of birth was disaggregated to three separate matching variables for the record linkage: day of birth (dd), month of birth (mm), and year of birth (yyyy). These three matching variables were then compared with the equivalent census variables in the record linkage. The date of birth for the NHI and NMDS Death Event File are independently collected, unless the decedent died out of hospital and had no previous hospitalisation. As such, having two independent sources of date of birth increases the discriminatory power of the record linkage.
Date of Birth (NMDS Death Event)	• matching variable • analysis of bias	The NMDS Death Event date of birth was used to generate the age for all decedents in the analysis of bias (i.e. differences between mortality records linked and those not linked). It was used for the analysis of bias to generate age at census night (in years) in preference to the NHI date of birth, as during the record linkage it appeared to be more accurate than the NHI DOB.
Sex (NHI)	• matching variable	As for date of birth, sex was available from both the NHI and NMDS Death Event File, usually independently sourced.
Sex (NMDS Death Event)	• matching variable • analysis of bias	-
Ethnic Group (NHI)	• matching variable • analysis of bias	As for date of birth, ethnic group was available from both the NHI and NMDS Death Event File, usually independently sourced. However, the manner of collection of the NHI and NMDS ethnic group variables also differed. The NHI ethnic group variable was only a single option (unlike the 1991 census multiple options), but it was supposed to be self-defined (i.e. the hospital admission clerk was supposed to ask the patient their self-identified ethnic group). Thus the NHI ethnic group was probably 'closer' to the census ethnic group variable than the NMDS one (see below) - therefore it was used for the analysis of bias in preference to the NMDS ethnic group. As a matching variable, it was classified in five hierarchical levels: Maori, Pacific, Asian, Other, European.
Ethnic Group (NMDS Death Event)	• matching variable	The NMDS ethnic group is collected by the undertaker. They are supposed to ask the family/whanau to 'self-identify' the decedents ethnic group, but this often does not happen. For this reason, enumeration of Maori deaths (and probably Pacific) is significantly less on the NMDS Death Event file than that if the NHI ethnic group is used. The NMDS Death Event ethnic group for 1991-94 was classified only as Maori, Pacific, and the Rest (the remainder), therefore it was specified as a matching variable with three values.
Country of Birth (NMDS Death Event)	• matching variable	Both the NMDS Death Event File and census include country of birth.
Meshblock (SNZ Vitals file)	• blocking variable	The meshblock code is not available directly from NZHIS, but was merged with the mortality data from the SNZ Vitals file, using a concatenated variable of Death Registration Office, Year, and Number. Meshblock is the most important blocking variable.
Vitals-CAU (SNZ Vitals file)	• blocking variable	The CAU including the meshblock from the SNZ Vitals file. If either the mortality or census record for a particular decedent had a miscoded meshblock, but miscoded to another meshblock in the same CAU, then blocking by the Vitals-CAU may result in a correct link. Also, about 10% of decedents had no meshblock, only a Vitals-CAU.

Table 2: Mortality variables used in the record linkage and analysis of bias

Variable	Purpose	Comments
Post-CAU (NMDS Health Event File)	• blocking variable	It was possible using the NMDS file to select a health event record (usually a hospitalisation) immediately after the 1991 census, but before the death event, for many decedents. This record had a domicile code for the usual residence at the time of that hospitalisation derived from NZHIS's own geocoding programme, and that domicile code directly corresponded with a SNZ CAU. If the decedent had changed their usual residence between census night and death, then blocking by Vitals-CAU would not have resulted in a link with a census record. It may be that the post-CAU variable gives the decedent's actual CAU on census night, enabling a correct link between the decedents mortality and census records.
Pre-CAU (NMDS Health Event File)	• blocking variable	As with the post-CAU variable, It was possible using the NMDS file to select the health event record immediately before the 1991 census for many decedents. If the decedent had changed their usual residence between census night and death, then it may be that the pre-CAU variable gives the decedent's actual CAU on census night, enabling a correct link between the decedents mortality and census records.
NHI-CAU (NHI File)	• blocking variable	The usual address on the NHI file may be different from that on the death registration form (particularly when the death occurred outside of a hospital), and any pre- or post CAU variables. Therefore, when a mortality record fails to link with a census record using other blocking variables, it may be that the NHI domicile code as a blocking variable gives the correct usual residence for the decedent on census night, enabling a link with a census record.
Occupation (NZSCO-90; NMDS Death Event)	• analysis of bias	The NZSEI occupational class was derived from the NMDS occupation variable. NZSEI occupational class was used to analyse the possible bias between mortality records linked to a census record, and those not linked.
Date of Death (NMDS Death Event)	• analysis of bias	The time between the census and death was used in the analysis of bias to investigate how record linkage rates changed over time.
Cause of Death (NMDS Death Event)	• analysis of bias	Much of the analysis of bias was for all deaths combined, but some were for separate groupings of cause of death.
Area Health Boards	• analysis of bias	Using the Vitals meshblock code, SNZ assigned area health board codes to the mortality records (USHBD01). This variable was then aggregated to regional health authority for use in the analysis of bias (i.e. analysis of bias of the record linkage by region).
Territorial Local Authority	• analysis of bias	Using the Vitals meshblock code, SNZ assigned TLA codes to the mortality records (USTLA01, 1989-base).
Urban Area	• analysis of bias	Using the Vitals meshblock code, SNZ assigned urban area codes to the mortality records (USUA01).

2.2.2 Census data

The 1991 census file is stored as a master file at SNZ. Only the variables necessary for this research were extracted to a smaller file. Subsets again of this file were used in the

record linkage, analysis of bias, and the final cohort analysis. For privacy reasons, and as prescribed under the Statistics Act (1975; see SNZ Security Statement page 19), only staff of SNZ have access to the census master file.

2.2.2.1 Geographic census variables used in the record linkage

The geographic census variables required for the record linkage were the usual residence meshblock and usual residence CAU. The former was used as a blocking variable in combination with the meshblock obtained from SNZ Vitals file for the mortality records. The census usual residence CAU was used as a blocking variable in combination with the available mortality CAU variables: Vitals-CAU, post-CAU, pre-CAU, and NHI-CAU. The census file meshblock and CAU codes are updated annually. However, the domicile codes on the mortality data were all in 1991-base. Therefore, all the census meshblock and CAU codes were back-coded to the 1991-base.

2.2.2.2 Personal census variables used in the record linkage

Complementary variables to the matching variables for the mortality data listed in Table 2 were created from census data. For both ethnic group and country of birth, this required aggregation of the primary census ethnic and country of birth variable values.

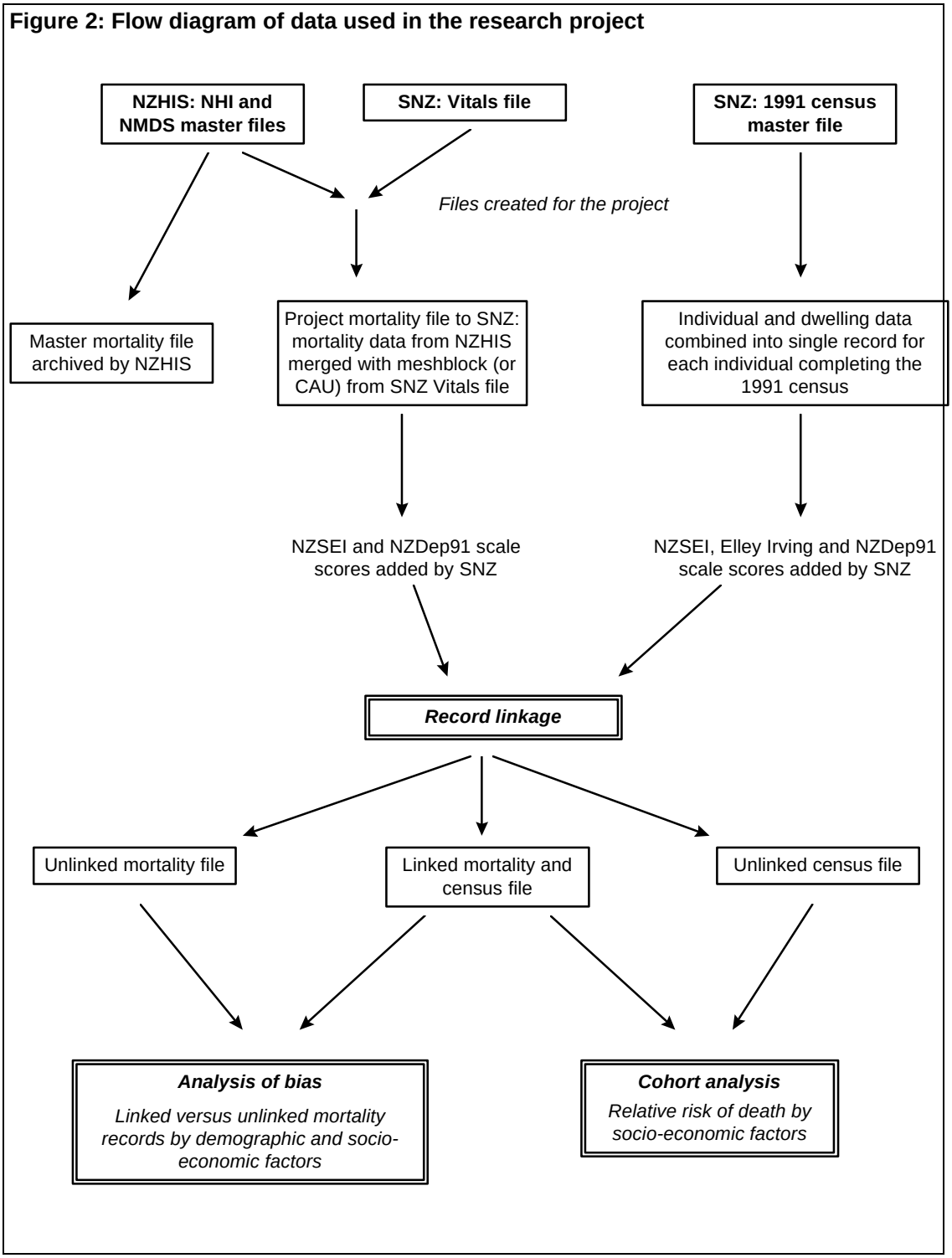
2.2.3 Flow of mortality and census data

The flow of mortality and census data is shown in Figure 2. The New Zealand Socio-Economic Index (NZSEI) and Elley Irving scale scores were assigned by SNZ. Only the NZSEI scale scores were assigned to the mortality data as the 1968 New Zealand Standard Classification of Occupations (NZSCO-68) code necessary to code the Elley Irving scale was not available for the mortality data. Both NZSCO90 (1990 codes; necessary for assigning NZSEI scores) and NZSCO68 codes were available for census data. After initial analyses of mortality data a problem with NZSCO90 codes on mortality data was unearthed – the NZSCO90 codes for deaths in 1991 (the first 10 months of deaths used in this project) were not necessarily correct. 1991 was a transition year for NZHIS between NZSCO68 and NZSCO90 codes. During 1991, the mortality records were actually first assigned a NZSCO68 code, and then this code was ‘converted’ to a NZSCO90 code (personal communication, Jim Fraser, Manager,

NZHIS, August 1999). It appeared that this conversion was not particularly good, and resulted in different NZSEI occupational class being assigned than if the written occupation had been directly coded to NZSCO90. Fortunately, NZSCO90 codes were not used in the record linkage. However, analyses of bias using occupational classes had to be redone discarding all decedents in the first year of follow-up.

Mortality and census files were then submitted to Automatch® for record linkage, and three output files obtained: an unlinked mortality file, an unlinked census file, and a linked file. The remainder of this Chapter details the methods used to conduct the record linkage, determine the accuracy of the record linkage, and assess the bias by demographic and socio-economic variables in the record linkage.

Figure 2: Flow diagram of data used in the research project



2.3 Record linkage strategy and match specifications

The process of record linkage is one of trial and error, and a combination of both science and art. The material that follows is thus just a consideration of general strategies - specific strategies that were ultimately used in the record linkage are presented as a component of Chapter 3: Results – Record Linkage.

2.3.1 Geocodes and pass order

A frequency distribution of the number of deaths by meshblock is shown in Table 3.

Table 3: Distribution of number of deaths per meshblock of usual residence for the 1991-94 mortality file

Number of deaths per meshblock	Absolute number of meshblocks	Percentage of meshblocks
0	16763	47.7%
1	8065	22.9%
2	4818	13.7%
3	2646	7.5%
4	1330	3.8%
5-9	1323	3.8%
10-14	129	0.4%
15-19	37	0.1%
20-29	30	0.1%
30-39	5	0.0%
40-49	5	0.0%
Total	35151	100%

As the most discriminating geographic variable, meshblock of usual residence was always used as the first blocking variable in any match-run strategy. The best ordering of the subsequent CAU passes was determined by trial and error.

2.3.2 Day, month and year of birth

Day, month, and year of birth were available from two data sets for the mortality records: the NMDS and the NHI file. These variables will not always agree between the NHI and NMDS file due to coding errors. Comparing these variables between the NMDS and NHI file, the following was noted:

- about five percent of NHI and NMDS days of birth did not agree: 1.4% disagreed by one day, 0.7% by two days, 0.4% by 3 days, 0.3% for each of 4 and 5 days, and then relatively little thereafter
- about two percent of NHI and NMDS months of birth did not agree: 0.8% by one month, 0.3% by two months, 0.2% by 3 months, and then relatively even thereafter
- about four percent of NHI and NMDS years of birth did not agree: 2.1% by one year, 0.4% by two years, 0.1% by 3 years, and then relatively flat thereafter.

Using the level of agreement/disagreement between the NHI and NMDS files as an approximation to the likely agreement/disagreement between mortality and census data, it seemed reasonable to allow for a tolerance or partial match on these variables. A decision was made to allow for a +/-1 tolerance for each of day, month, and year of birth by using the PRORATED command in Automatch®. This command assigns the full agreement weight to an exact agreement, a weight midway between the full agreement and full disagreement weight for a partial agreement (ie. +/-1), and the full disagreement if the disagreement is by two or more.

2.3.3 Ethnic group

Ethnic group was also recorded on both the NHI and NMDS file. For the 1991-94 period, the NMDS file could only be categorised usefully as Maori, Pacific people, and the 'Rest'. The NHI file could be categorised as Maori, Pacific, Asian, Other and European. Therefore, the census ethnic group data was categorised likewise to give two separate ethnic variables: a three level variable for matching with the NMDS ethnic group variable, and a five level ethnic group variable for matching with the NHI ethnic group variable.

As the census allows for multiple ethnic groups, the census hierarchical categorisation of ethnic group was used to categorise all census records to just one ethnic group. This immediately introduced a bias: even if the single ethnic group coding for health data was ideal in that it always assigned individuals to their 'first choice' sole ethnic group, the census records would be assigned Maori even if the individuals 'first choice' ethnic group was European. This bias added to the known discrepancy between health and

census data where Maori (and Pacific) are more likely to be coded on census data than health data for the same individual.[25]

A further problem that became apparent during the record linkage was that many decedents were assigned as 'Other' (code = 54) on the NHI file when they were actually 'Other European' (eg. non-New Zealand European). These people would have been coded as European by census data. Given this discrepancy between health and census data, it would have been ideal to specify Automatch® not to assign a disagreement weight, but that was not possible. (It may be better in any future record linkage to assign the NHI ethnic group to the same three levels as the NMDS ethnic group variable.)

2.3.4 Occupation

Occupation is recorded on both the NMDS Death Event file and the census file, but we decided not to use it as a matching variable for three reasons. First, over 80% of women had no occupation stated in the NZHIS mortality data. Therefore, occupation was likely to be more successful for matching male records, and frequency ratios would require calculation separately by sex. Second, occupation was ascertained differently on the death registration form (usual occupation) compared to the census (main occupation in last four weeks), although both use the New Zealand Standard Classification of Occupations (NZSCO90). Third, using occupation as a matching variable may bias the record linkage by socio-economic status, as people with stable careers (associated with higher socio-economic status) would be more likely to be linked than people with changing or no occupations. (As noted above, it was fortunate that occupation was not used as a matching variable for a further reason – the NZSCO90 codes on 1991 mortality data were not particularly accurate.)

Occupation was, however, useful during determination of match cut-offs and clerical review rules by assisting identification of likely true links.

2.3.5 Value specific *m* probabilities

For ethnic group and country of birth the *m* probability varied by variable value. For example, New Zealand as a birthplace would have a higher *m* probability than America. To not allow for these value specific *m* probabilities would mean that a person born in New Zealand (and recorded as such on both census and NMDS data) would receive a smaller agreement weight than was 'correct'. Therefore, the Automatch® command MPROB was used to assign value specific *m* probabilities.

2.3.6 Maximising the benefit of NMDS and NHI file measures for the same variables: to array, or not to array?

Automatch® allows variables to be specified as an array. For example, the NMDS and NHI day of birth variable could be specified as an array for matching with the census day of birth variable. This would result in:

- a full agreement weight if both the NMDS and NHI day of birth variable agreed with the census variable
- a full disagreement weight if both the NMDS and NHI day of birth variable disagreed with the census variable
- half the agreement weight if either the NMDS or NHI day of birth variable agreed with the census variable.

By way of comparison, specifying the NMDS and NHI day of birth variable separately would result in twice the agreement weight if they both agreed, twice the disagreement weight if they both disagreed, and a combined weight midway between the agreement and disagreement if one agreed but the other did not. The latter midway weight is the key difference between using an array and just specifying the variables separately: an array would assign half the agreement weight, a weight that would usually be considerably more than the sum of the agreement and disagreement weight. In effect, an array allows for the fact that when the NMDS and NHI variables disagree, one of the NMDS or NHI variables is certainly wrong *but* one is almost certainly correct - they are not separate independent variables as suggested by specifying the variables separately. An array specification of day, month and year of birth was therefore preferred. However, Automatch® was unable to compute simultaneously: a tolerance

of +/-1; value specific *m* probabilities; and an array for any particular matching variable. All were considered important contributors to the discriminating power of the record linkage. A decision was made to dispense with the arrays, retain a +/-1 tolerance for day, month, and year of birth, and calculate value specific *m* probabilities. The record linkage would remain balanced in that all of sex, day, month, and year of birth, and ethnic group were recorded on both the NHI and NMDS file, and hence would be 'counted twice'. The only exception was country of birth, which was recorded on the NMDS file only.

Consideration was given to including the country of birth variable as two separate (but identical) variables in the record linkage to completely 'balance' the probabilistic record linkage, but was decided against for the following reasons. First, in the majority of instances country of birth would add very little discrimination as most decedents (80%) were born in New Zealand. Second, and related, there were concerns regarding the systematic errors in the recording of country of birth. For example, if no country of birth was recorded on the BDM28 (death registration form) then New Zealand was entered as the default. Whilst an agreement of New Zealand on both the census and mortality record would attract very little positive weight, a disagreement of New Zealand and another country may attract a sizable negative weight. Therefore, including country of birth only once as a matching variable was considered a conservative measure against the understood deficiencies of country of birth as a matching variable.

2.4 Determining the accuracy of the record linkage

There was no gold standard against which to determine the sensitivity and specificity of the probabilistic record linkage in this research. For example, names and addresses were not available on a subset of both mortality and census records, against which the anonymous record linkage could be validated. Therefore, the quality of the record linkage could only be estimated by indirect measures. These included:

- the percentage of mortality records linked to a census record at each pass
- the total percentage of mortality records linked to a census record for the full match-run.

The latter, the total percentage, *approximates* the sensitivity of the record linkage ([number of 'true' links detected]/[total number of 'true' links among all possible individual comparisons of mortality and census records]), assuming:

- the number of false positive links are negligible compared to the total number of accepted links (numerator bias)
- the number of decedents without a census record is negligible (denominator bias; some decedents may have been overseas on census night, or simply not have completed the census).

There was no similar, and simple, approximation for the specificity of the record linkage. The specificity would be: [the number of 'incorrect' links rejected]/[total number of 'incorrect' links]. There is a method for estimating the number of false positive links ("incorrect" links) using the probabilistic weights, as described by Newcombe.[22 26] We will refer to this method as the 'absolute weight method'. The absolute weight method is prone to bias from correlated errors between matching variables. Thus, alternative methods of estimating the number of false positive links were developed specifically for this project, relying more on the empirical nature of the data rather than the theoretical distribution of the probabilistic weights. These methods developed specifically for this project are applicable only to record linkage projects were there could be only one true link for each record, so-called 'best linkage'. This was the case with linkage of census and mortality records, as each individual only

completes one census form and can only die once. The first method we present specifically developed for this project, the 'chance method', simply estimates the number of false positive links where the records agree exactly on the matching variables. The second method, the 'duplicate method', estimates the number of false positive links by using the observed number of mortality records with zero, one, or two census record links. This duplicate method requires that the probability of a false link purely by chance is constant for all mortality records, and uses binomial combinatorial probabilities to estimate the number of false positive links.

2.4.1 The absolute weight method for estimating the number of false positive links

The weight assigned by Automatch® is the log (base 2) of the relative odds.[23 27] The relative odds does not inform how likely a link is to be a true link, just its relative likelihood compared to other links. But the relative odds can be transformed to the absolute odds, where the absolute odds is the betting odds for any single link being a true link. To make this transformation, the following information is required:[22 26 28 29]

- the number of file A and file B records processed in each pass, where file A is the initiating file and file B is the searched file
- the number of file A records with a true link somewhere in the file being searched (initially estimated from sample linkages, and later refined).

The formula for the absolute total weight (W*; complementary to the absolute odds) is:

$$W^* = W + \log_2 [N(A \mid t) / N(\text{file A})] + \log_2 [1/ N(\text{file B})]$$

where:

W = the relative total weight
N(A | t) = number of records in file A with a true link in file B
N(file A) = number of records in file A
N(file B) = number of records in file B

This formula assumes that there is no blocking. To allow for blocking, the formula can be adjusted for the component weight that the blocking variable would have added if it had been included as a matching variable rather than a blocking variable. Alternatively, the formula can be applied individually to each block, or the average for each block. Applying the formula to each block individually would be subject to instability from small block sizes, and computationally onerous. Using the pragmatic option of the 'average block' introduces possible bias from variation in block size and the relative distribution of file A and file B records by block.

The average block method was used in this record linkage project. An extra feature of this research allowed further simplification of the formula: each mortality and census record could only be linked once (so-called 'best linkage'). Therefore, the distinction between the initiating (file A) and search file (file B) is not necessary. The formula used in this research was:

$$W^* = W + \log_2 \frac{\text{Av(t links in block)}}{\text{Av(mortality)} \times \text{Av(census)}}$$

where:

Av(t links in block)	= the average number of true links for each block
Av(mortality)	= the average number of mortality records in each block
Av(census)	= the average number of census records in each block .

The absolute odds can be calculated by taking the anti-log (base 2) of W*. The absolute odds can then be transformed to the positive predictive value (PPV= absolute odds/[1 + absolute odds]).

The weight assigned by Automatch® in this research was not equivalent to W in the above formula, but instead was approximately twice W. That was because each of the matching variables (except country of birth) were used twice, once for the NHI file and once for the NMDS file. (The distortion from using country of birth once only as a matching variable will be negligible in most instances, as for the majority of decedents

(New Zealand born) it contributes little weight). Therefore, the weight assigned by Automatch® had to be halved first before being introduced to the above formula.

The absolute weight method is prone to bias, including:[22 26]

- correlated disagreements between matching variables for a true link (eg. month of birth may be more likely to disagree for a true link when day of birth also disagrees)
- correlated agreements between matching variables for a true link (ie. the reverse of the above)
- age-related bias due to the alteration in prior probability of death for any cohort followed over time.[26]

The first two sources of bias may be a problem in this research. Additionally, there will be correlated agreements and disagreements between the NMDS and NHI file variables that would be better represented by arrays, but, as already discussed (page 42), we decided not to use arrays given the use of partial agreements (PRORATED) and value specific *m* probabilities (MPROB).

The last source of bias, age-related bias, is not a major factor in this project as the follow-up was only for three years: there would be little change in the prior probability of death by year for each member of the New Zealand population over a three year period.

Finally, bias and the methods used for record linkage in this research aside, the absolute weight method (as specified by Newcombe and others) was not equivalent to the alternative methods we developed for estimating the false positives. This was because the absolute weight method counts any duplicate links (e.g. one mortality record matching with two census records, or what Automatch® labels a ‘DA pair’) as false positives, in addition to those undetected false positive single links. We discarded all duplicate pairs, given that there can only be one link between each census and mortality record. Therefore, when comparing the absolute weight method with the other methods to estimate false positives in this project, we first adjusted for duplicate pairs.

2.4.2 The chance method for estimating the number of false positive links

For any given comparison of a mortality record with a census record, there is a probability that they will agree *exactly* on all matching variables despite not being the same individual's census and mortality record. This probability of an exact match purely by chance can be calculated using the *u* probabilities.

For a New Zealand born, New Zealand European, the probability of an exact link due to chance alone is about:

$$0.5\times0.033\times0.083\times0.013\times0.80\times0.80=0.0000114, \text{ or } 1.14\times10^{-5}$$

where the *u* probabilities are for sex (1/2 = 0.5), day of birth (1/30 = 0.033), month of birth (1/12 = 0.083), year of birth ('average' 1/75 = 0.013), ethnic group and country of birth (both 'average' probabilities of about 0.8), respectively. The *u* probabilities for a Maori born in New Zealand would suggest an even smaller probability of a chance link (approximately 1.7×10^{-6}), but the decrease would probably not be this great due to residential clustering by ethnic group. The value of 0.013 for age is only an 'average' value, and for older decedents the *u* probability will be nearer 0.006 (0.006 is the proportion of people aged, say, 70 years in New Zealand). Whilst there is also residential clustering by age, it is probably not enough to offset the effect on chance links from reduced *u* probabilities for older people. As most deaths are among older people, this would suggest that the approximate 1.14×10^{-5} estimate is conservative. Considering the above, a figure of 1×10^{-5} seems a reasonable best estimate of the probability of any single mortality record agreeing exactly with any single randomly selected census record, on the six matching variables above.

2.4.2.1 Estimating the number of false positive links using the chance method

Among any set of observed exact links between census and mortality records in a given **meshblock**, there may be false positive links, including:

- false positive links involving a mortality record that also has a true exact link with a census record in the given meshblock

- false positive links involving a census record that also has a true exact link with a mortality record in the given meshblock
- false positive links involving a mortality record and a census record where neither record has an exact link within the given meshblock.

This distinction is important. The first type of false positive link involves a mortality record linked to two (or, rarely, more) census records. Automatch® will label one link as an MP pair ('match pair') and the other as a DA pair ('duplicate A file pair' where the 'A file' refers to the census file). When both the DA and MP pair are exact matches they have the same weight-score, and it is impossible to tell which link is more likely to be the 'true' link (the MP and DA labels are assigned at random). In this project, these false positive links were avoided by simply discarding both links (i.e. both the DA and MP pair) as there was only a 50:50 odds of selecting the true link. That is, sensitivity was sacrificed to minimise the number of false positives accrued during the record linkage.

The second type of false positive link was the reverse of the first - it involved a census record that was linked to two (or, very rarely, more) mortality records. Automatch® will label one link as a DB pair (where 'B' refers to the mortality file), and the other as an MP pair. As above, these false positive links were avoided by simply discarding both links. Note that this type of false positive link was much less common than the first false positive link in this project due to the large difference in size between the mortality and census files.

The exactly matching false positive links in the third bullet point were those that would be not be detected in this project. The chance method attempts to estimate the number of these false positive links, as follows.

To estimate the probability of a given mortality record having a purely chance match with a census record, the binomial distribution can be used:

$$ {}_nC_r\,(1-p)^{(n-1)}\,p^{(r)} $$

where:

n is the number of trials (comparisons with different census records);
r is the number of successes (links purely by chance);
${}_nC_r$ is the binomial coefficient [n!/(r!(n-r)!)];
p is the probability of 'success'.

What should 'r' be? No successes (r=0) means that no census records are linked to the given mortality record, and hence no false positive links are possible. Two or more successes ($r \geq 2$) means that the given mortality record is linked to two or more census records purely by chance. As two or more purely chance exact links for one mortality record would have the same weight score, they would be detected and discarded. Thus, we were interested only in single (r=1) links purely by chance for each mortality record.

'p' is the probability of an exact match (purely by chance) for any given comparison of one mortality with one census record. This was estimated above as approximately 1×10^{-5}. The number of trials, n, is the number of census records that each mortality record is compared to within the meshblock. The average meshblock size was approximately 100. The number of trials, though, for any given mortality record *with no true exact census record link* in a meshblock will possibly be less – any true link(s) in the same meshblock for *other* mortality record(s) will correspondingly reduce the number of census records available for comparison. In this project there was approximately one true link per meshblock (approximately 30,000 true links spread over approximately 30,000 meshblocks). As on average this will have a negligible effect on n (100 – 1 = 99), we simply assumed that n=100.

Recapping thus far, the above binomial expression reduces to:

$${}_{100}C_1 \times (1\text{-}0.0001)^{(100-1)} \times 0.00001^{(1)}$$

$$= 100 \times (0.99999)^{99} \times 0.00001 = 0.000999 \approx 0.001$$

That is, for any single mortality record with no true exact census link in a given meshblock, there was an expectation of 0.001 false positive links occurring purely by chance. (Note the above expectation is approximated by 100 (average meshblock size) $\times$ 0.00001, without needing the binomial coefficients, as n was relatively large and p

was relatively small.) As the expectation of a single false positive link for each mortality record is much less than one (0.001), it is not inappropriate to refer to 0.001 as the probability of any one mortality record with no true exact link having a single false positive link.

The final step was to estimate the actual number of exact matching false positive links. Strictly, this would be the number of mortality records with no true exact census link in the given meshblock, multiplied by the 0.001 probability. However, the former set of mortality records is unobservable – most will be observed as unlinked mortality records, but some will be false positive links indistinguishable from the true links (i.e. those we are trying to estimate). The number of false positive links can be estimated by iteration, though. In the first iteration, the number of mortality records with no observed link is multiplied by 0.001 to derive a crude estimate of the number of false positives. In the second and subsequent iterations, the number of mortality records with no true exact census link is estimated more accurately by using the number of mortality records with no observed link, *plus* the estimated number of false positives from the previous iteration. As 0.001 is a small probability, the number of estimated false positives from the first iteration changes little with subsequent interations. For example, assume that 10,000 mortality records had no observed link with a census record. The first iteration estimates $10{,}000 \times 0.001 = 10$ false positives. The second iteration estimates $10{,}010 \times 0.001 = 10.01$ false positives. The third iteration estimates $10{,}010.01 \times 0.001 = 10.01001$ false positives, and so on.

The above workings apply to the first pass by **meshblock** in the record linkage. An assumption was made that the average meshblock size was 100. As it transpired, the average meshblock size *among meshblocks that actually had a link* was 132 – a consequence of a link being more likely with increasing meshblock size. However, the above workings are still valid. For the record linkage using **census area units** as the blocking variable, month of birth was also used as a blocking variable – other wise there was ‘block overflow’ with too many comparisons with some blocks for Automatch® to process. The effect on the chance method at the CAU-level was twofold. First, the underlying probability of a false positive link between any one

mortality record and any one census record increased by a factor of 12 to allow for the loss of month as a matching variable. Second, the average block size was approximately 200.

2.4.2.2 The effect of variation in the meshblock or CAU size

What would be the effect on the chance method of varying the block size? Assume that there were 100 deaths, 50 meshblocks of size 50, and 50 meshblocks of size 150. Assuming that deaths are distributed relative to meshblock size, the expected number of deaths in meshblocks of size 50 will be $100\times[50\times50]/([50\times50]+[50\times150]) = 25$. The remaining 75 deaths will be in meshblocks of size 150. Assume that 50% of mortality records had a true exact link with a census record and this probability was constant across meshblocks, and that p was 0.00001 as above. Using the method described above, and summing over the 100 meshblocks, the expected number of false positive links purely by chance is approximated by the sum of:

$$\{(75\times0.5)\times150\times0.99999^{149}\times0.00001\} + \{(25\times0.5)\times50\times0.99999^{49}\times0.00001\}$$
$$= 0.056 + 0.006$$
$$= 0.062$$

where 75×0.5 and 25×0.5 are the expected number of mortality records without a true link in meshblocks of size 150 and 50, respectively. If the meshblocks had all been the same size (n=100), 0.05 false positives would have been expected if the meshblocks were all of size 50. That is, in this example the expected number of false positives was 25% (1-(0.062/0.05)) greater when the meshblocks were unevenly sized. More generally, unequal meshblock size will result in a higher number of expected false positives than if all meshblocks were of equal size.

Given that there is variation in meshblock and CAU size, it is desirable to allow for their effects in the chance method - but without making the method too complex. As the average meshblock (or CAU) size was readily available, but not the full distribution of meshblock sizes, the simplest way to allow for variation in block size was to inflate the underlying 0.00001 estimate of the probability of any one mortality and any one census record matching exactly by chance, and apply it directly to the average (mean)

block size. A probability of 0.000012 was considered appropriate for application to average meshblock or CAU sizes.

2.4.2.3 Possible inaccuracies in the chance method

The chance method was subject to two main sources of error. First, the chance method is dependent on the threshold weight selected for exact (or near exact) links. There is no clear demarcation above which all links are exact matches, and below which all links are non-exact matches. This is a function of the probabilistic record linkage where a link with uncommon variable values and one minor disagreement can receive a higher weight than a link with common variable values yet exact agreement on all matching variables.

Second, the underlying probability estimate of 1.2×10^{-5} for any one mortality record to agree with any one census record was subject to many assumptions, and may have been inaccurate.

2.4.3 Duplicate method for estimating the number of false positive links

The above chance method applies to exact (or 'near-exact') links only. In this section a method for estimating the number of false positive links is presented that is applicable to non-exact (but still highly likely) *and* exact links. The duplicate method involves simultaneously solving the combinatorial probabilities for zero, one, or two census links for a given mortality record. Assume that above a given probabilistic weight-score, there is a uniform probability, 'p', that any one mortality record will have a purely chance link with any one census record. Let 't' be the probability that a mortality record has a true link, and 'n' be the number of census records (trials) compared to each mortality record. Thus:

(1) Pr (no true link and 0 false links)	= [1-t]	[		$(1-p)^{n}$]
(3) Pr (no true link and 1 false link)	= [1-t]	[n	p	$(1-p)^{n-1}$]
(5) Pr (no true link and 2 false links)	= [1-t]	[n(n-1) /2)	p^2	$(1-p)^{n-2}$]
(7) Pr (no true link and 3 false links)	= [1-t]	[n(n-1) (n-2) / 6)	p^3	$(1-p)^{n-3}$] etc.
(2) Pr (1 true link and 0 false links)	= [t]	[		$(1-p)^{n-1}$]
(4) Pr (1 true link and 1 false link)	= [t]	[(n-1)	p	$(1-p)^{n-2}$]
(6) Pr (1 true link and 2 false links) etc.	= [t]	[(n-1)(n-2) / 2)	p^2	$(1-p)^{n-3}$]

Note that the sum of the odd-numbered sequence is just (1-t) since the terms in the second brackets are the binomial probabilities of observing 0, 1, 2, ... n false links in n comparisons and thus sum to unity. Similarly, the even-numbered sequence sums to t. Thus the sum of all possible probabilities is (1- t) + t = 1.

In practice, at and above a given weight-score, we observe the proportion of mortality records with zero, one, two, etc., census record links. Let:

X = the proportion of unlinked mortality records	= (1)
Y = the proportion of singly linked mortality records	= (2) + (3)
Z = the proportion of duplicate linked mortality records	= (4) + (5)

where (1), (2), …, (5) refer to the numbered equations above on page 54; and the proportions apply to a specified weight cut-off in the linkage. These proportions (X, Y, Z) estimate the underlying probabilities of an unlinked mortality record, a singly linked mortality record, and a duplicate linked mortality record, and thus equal the sums of equations (1), (2) and (3), and (4) and (5) from page 54. Multiplying the equation for Y by (n-1) (1-(1-p)) / (1-p), subtracting the equation for Z, and then substituting X / $(1-p)^{n}$ for (1-t) (from the equation for X), we get a quadratic in (1-p):

$$[\, n(n-1)X + 2(n-1)Y + 2Z \,] (1-p)^2 - [\, 2n(n-1)X + 2(n-1)Y \,] (1-p) + [\, n(n-1)X \,] = 0$$

where n is the average number of census records in a block – approximately 100 for a meshblock pass. The equation has two roots. Back substitution gives values for p and t. The correct one of these two roots will give $t < 1$ and $(1-p) + p = 1$. As a further check, the values of (1-p), p, and t must also satisfy equation (2) on page 54.

Assuming that when there is a true census record link it is the highest scoring link for each mortality record, the proportion of all mortality records involved in false positive links is derived from the corresponding probabilities expressed in the equations (3), (5), (7), etc., above. (Equations (5), (7), etc., also contribute false positive links as the highest weight-scoring pair for each mortality record with two, three, or more false positive links would be accepted.) These equations can be solved by back substituting the derived values for p and t.

Consider an example of 10,000 mortality records submitted to a meshblock pass with a given weight cut-off above which links were accepted and below which links were rejected. Assume that for this weight cut-off, there were 4,500 mortality records with no census links, 5,400 mortality records with just one census link, and 95 mortality records with two census links. (The 5 remaining mortality records have three or more census record links). Solving the quadratic equation, the value of p is 0.00017, i.e. each comparison of one mortality record with one census record is assumed to have a 0.00017 probability of having a weight-score above the cut-off. Also, t is given as 0.542. Back substitution of p and t gives a probability for equation (3) (the probability of no true links and one false positive link) of 0.00765, corresponding to 76.5 (i.e. $0.00765 \times 10{,}000$) expected false positive links. The probability for equation (5) (the probability of no true links and two false positive links) is 0.00006, corresponding to 0.6 expected false positive links. The contribution of higher odd number equations is negligible, thus the expected number of false positive links is 77 in this example.

2.4.3.1 Possible biases in the duplicate method

Assuming that the highest scoring duplicate link is the true-link may not always be correct.

As with the chance method, the duplicate method assumes that false positive links arise only for mortality records with no true census link. Occasionally a false-positive link may have a higher weight score than the true-link involving the same mortality record (due to miscoding of either the census or mortality record for the true link). This

scenario would present as different weight scoring duplicate links for a given mortality record, and accepting the highest weight-scoring link would incur a false positive link. However, such a scenario should be uncommon – if the true-link is miscoded on either files, it is still more than likely that it would be the highest weight scoring link compared to any duplicate links. Calle et al (1993) provide empirical evidence for this assertion.[18] Using a probabilistic record linkage methodology in a project very similar to the NZCMS, they incurred no false positive links when accepting the highest weight-scoring link of duplicate pairs. The project reported by Calle et al (1993) used more discriminatory matching variables than were available in the NZCMS – thus it is prudent to conclude that we may incur some false positive links under this scenario, but probably not many.

Simply using the number of DA pairs in the Automatch® output may be misleading

Routine reports in Automatch® give the number of MP and DA pairs above a given weight cut-off. But these routine reports do not give the number of mortality records associated with two, three of more census links. If the number of mortality records with four or more census links is assumed to be negligible (i.e. zero), then the number of mortality records with two and three census record links can be deduced from the total number of mortality records submitted to the pass, the number of MP pairs, the number of DA pairs, and the number of residual (unlinked) mortality records for a given pass. However, during the actual development of a record linkage match-run, time may be limited and the number of DA pairs on the Automatch® simply taken as the number of mortality records with a duplicate link. What is the effect of this short-cut?

Returning again to the example on page 55, the number of observed DA pairs would have been 105 (95 DA pairs for the 95 mortality records with two census links, and 10 DA pairs for the 5 mortality records with three census links). If just the proportion of MP pairs (still 0.54), the proportion of DA pairs to all mortality records submitted (0.0105), and the quickly deduced residual proportion of unlinked mortality records (1 – 0.54 – 0.0105 = 0.4495) were used in the quadratic equation for (1-p), then the expected number of false positive links would be 91. This estimate is 18% greater than

the previous estimate of 77. Thus, the actual number of *mortality records* involved in duplicate and triplicate pairs should be estimated, rather than just using the number of DA pairs in the Automatch® output as an estimate of the number of mortality records with two census links.

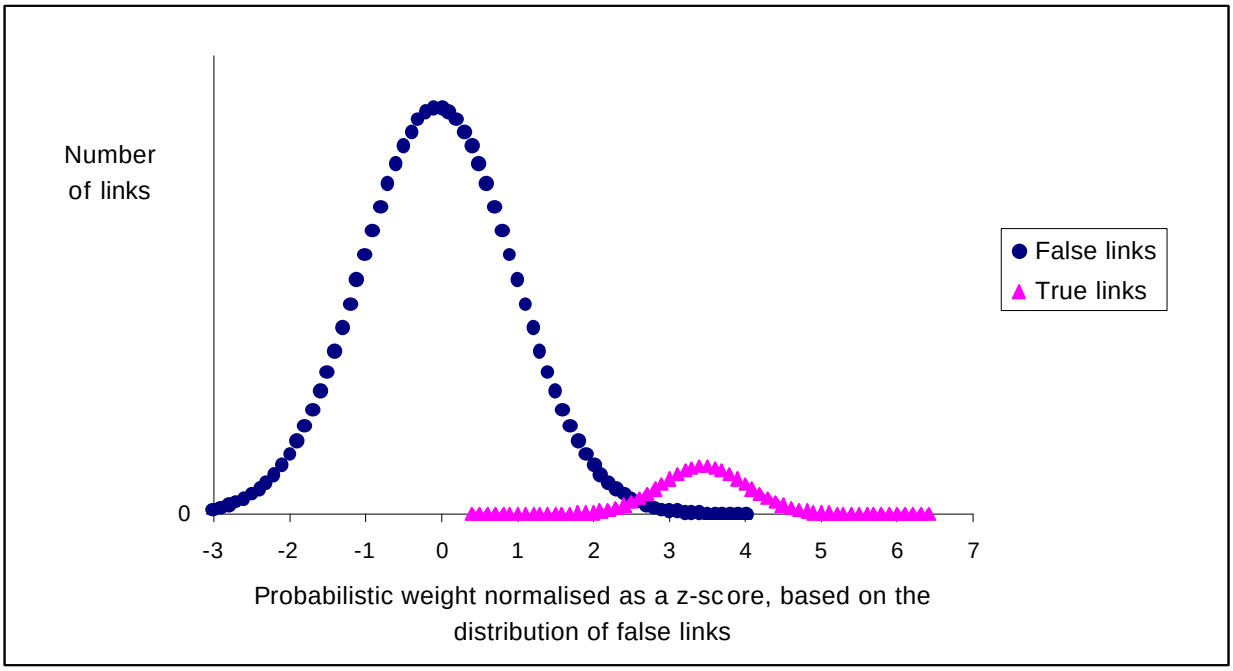
What is the effect of 'p' not being constant

For the binomial coefficients and distribution to be valid, the 'p' in the expression:

$${}_nC_r\,(1-p)^{(n-1)}\,p^{(r)}$$

must be the same for every single 'trial'. This is not the case with probabilistic record linkage. Figure 3 shows a 'typical' distribution of false links and true links by weight score for a probabilistic record linkage project, where the weights are converted to a z-score for the distribution of false links. (The distribution of both false links and true links are approximately normal by the weight score.[22])

Figure 3: Distribution of false and true links by weight score (normalised as z-score based on distribution of false links) for a 'typical' probabilistic record linkage



The general task of record linkage is to set the cut-off at the 'best' place between the two peaks in Figure 3 that maximises the number of true links accepted, but minimises

the number of false links accepted. For the purposes of this discussion, focus on the right hand tail of the false link distribution as it crosses the true-link distribution – for any cut-off used in the record linkage the false links will arise from the tail of the false-link distribution to the right of the cut-off. As the z-score increases, the number of false links decreases in an exponential manner. The underlying probability of any one mortality record having a link with any one census record at a given z-score (or probabilistic weight) will also decrease in a similar manner. The duplicate method implicitly assumes 'p' is constant above any given weight cut-off. What is the effect of this assumption?

To conduct a sensitivity analysis of the likely effect of this assumption, we constructed a distribution of p that varied proportional to the z-score of the distribution of false links. That is the underlying probability of any one census and any one mortality record linking with a probabilistic weight score in a given z-score range is proportional to the Normal distribution of false links. This is presented Table 4. The first column simply denotes the z-score range. The second column is the proportion of the total area under the Normal distribution that is included in each 0.1 unit z-score range. Thus the total of this column (0.00621) is simply the proportion of the Normal distribution that has a z-score > 2.5. The third column, w_i, is the proportion of the z-score distribution > 2.5 that is within each 0.1 z-score range. For example, for 2.5-2.6 it is equal to 0.00155/0.00621 = 0.2492. Correspondingly, the sum of this third column is one. The fourth column gives p_i (the z-score stratum value for p, the underlying probability of any one mortality record and any one census record having a false link at that stratum of probabilistic weight score) when the overall p for z > 2.5 is 0.002. To calculate p_i, we have assumed that:

$$p_i \propto w_i$$

Therefore:

$$p_i = c \times w_i$$

where c is some constant. Further, note that:

$$\begin{aligned} Av(p) &= 0.002 \\ &= \sum_i (p_i \times w_i) \\ &= \sum_i (c \times w_i^2) \\ &= c \times \sum_i (w_i^2) \end{aligned}$$

and therefore:

$$c = Av(p) / \sum_i (w_i^2) = 0.002 / \sum_i (w_i^2)$$

For the example in Table 4 the sum of w_i^2 is 0.1479. Thus c equals 0.01352. Using this value of c, p_i equals $0.01352 \times w_i$ for each stratum. For example, for the z-score range 2.5 to 2.6 p_i equals $0.01352 \times 0.2492 = 0.00337$. As a check of these calculations of p_i, the sum of $p_i \times w_i$ should equal 0.002 (Av(p)) – it does for the data presented in Table 4.

Table 4: Workings for the sensitivity analysis of p varying by probabilistic weight (modeled as the normalized z-score of the distribution of false links) when: z > 2.5; the average probability of any one mortality record having a false link with any one census record above z=2.5 is 0.002; and a meshblock pass of average size n=100 is assumed

z-score range	Marginal Normal distribution	Marginal Normal distribution when z>2.5 w_i	Stratum p when av(p) = 0.002 p_i	Weighted probability of 1 false link $w_i \times {}^{100}C_1 \times p_i \times (1\text{-}p)^{99}$	Weighted probability of 2 false links $w_i \times {}^{100}C_1 \times p_i^2 \times (1\text{-}p)^{98}$	Weighted probability of 3 false links $w_i \times {}^{100}C_1 \times p_i^3 \times (1\text{-}p)^{97}$
2.5-	1.55E-03	0.2492	0.00337	5.81E-02	9.23E-03	9.68E-04
2.6-	1.19E-03	0.1924	0.00260	3.72E-02	4.56E-03	3.69E-04
2.7-	9.12E-04	0.1469	0.00199	2.30E-02	2.15E-03	1.33E-04
2.8-	6.89E-04	0.1110	0.00150	1.38E-02	9.72E-04	4.54E-05
2.9-	5.16E-04	0.0831	0.00112	7.98E-03	4.22E-04	1.47E-05
3.0-	3.82E-04	0.0616	0.00083	4.50E-03	1.77E-04	4.56E-06
3.1-	2.80E-04	0.0452	0.00061	2.48E-03	7.11E-05	1.35E-06
3.2-	2.04E-04	0.0328	0.00044	1.33E-03	2.77E-05	3.81E-07
3.3-	1.46E-04	0.0236	0.00032	6.94E-04	1.04E-05	1.03E-07
3.4-	1.04E-04	0.0168	0.00023	3.55E-04	3.79E-06	2.67E-08
3.5-	7.35E-05	0.0118	0.00016	1.77E-04	1.34E-06	6.64E-09
3.6-	5.13E-05	0.0083	0.00011	8.68E-05	4.56E-07	1.58E-09
3.7-	3.55E-05	0.0057	0.00008	4.16E-05	1.51E-07	3.62E-10
3.8-	2.43E-05	0.0039	0.00005	1.95E-05	4.85E-08	7.94E-11
3.9-	1.64E-05	0.0026	0.00004	8.96E-06	1.51E-08	1.67E-11
4.0-	1.10E-05	0.0018	0.00002	4.03E-06	4.55E-09	3.39E-12
4.1-	7.31E-06	0.0012	0.00002	1.78E-06	1.33E-09	6.59E-13
4.2-	4.81E-06	0.0008	0.00001	7.69E-07	3.79E-10	1.23E-13
4.3-	3.13E-06	0.0005	0.00001	3.26E-07	1.04E-10	2.21E-14
4.4-	2.01E-06	0.0003	0.00000	1.35E-07	2.79E-11	3.80E-15
Total	0.00621	1		0.1556	0.0193	0.0018

The last three columns of Table 4 are the weighted contributions of one, two, and three false links at each stratum to the overall probability that any one mortality record will have at least one false link above a z-score of 2.5. Summing the last three column totals (0.1556 + 0.0193 + 0.0018) gives 0.177 – the overall probability in the given scenario of one mortality record having a false link in a meshblock of average size 100. How does this 0.177 probability compare with that that would be estimated if we simply used the average p (0.002) that would be determined by the duplicate method? Ignoring the stratum specific p_i in Table 4, the estimated overall probability would be:

$$\{100 \times 0.002 \times (1 - 0.002)^{99}\} +$$
$$\{(100 \times 99 / 2) \times 0.002^2 \times (1 - 0.002)^{98}\} +$$
$$\{(100 \times 99 \times 98 / 6) \times 0.002^3 \times (1 - 0.002)^{97}\}$$

$$= 0.181$$

which is little different from the stratum specific value of 0.177. Thus it appears from the sensitivity analysis for the scenario described in Table 4 that the error from just using the average value of p (calculated by the duplicate method) compared to using some plausible model of stratum specific p_i is negligible.

Table 5 presents a sensitivity analysis at the meshblock-level for varying values of average p. The second to last row of Table 5 summarises the results from Table 4 for an average p of 0.002. The results in Table 5 show that for all the specified values of average p there was little difference between the duplicate and stratum specific methods. The values of p in Table 5 included the range of values of p actually encountered in the record linkage. Furthermore, sensitivity analyses at the CAU-level (not presented here) demonstrated the same conclusion – within the range of average values for p encountered in this project, there did not appear to be any substantial bias for the duplicate method due to actual systematic variation in p by weight-score.

Table 5: Probability of a false link for each mortality record as calculated by the stratum specific and duplicate methods at the meshblock-level: sensitivity analysis for varying average underlying probability (p) of any one mortality record having a false link with any one census record above a given cut-off

Average p	Probability of a false link anywhere *above* the cut-off weight		
	Stratum specific method	Duplicate method	% difference
0.00005	0.005	0.005	0%
0.00010	0.010	0.010	0%
0.00020	0.020	0.020	0%
0.00050	0.048	0.049	1%
0.00100	0.094	0.095	1%
0.00200	0.177	0.181	3%
0.00500	0.369	0.393	7%

What is the effect of varying block size?

For the chance method, we concluded in Section 2.4.2.2 (page 52) that variation in the average block size caused the chance method to underestimate the number of false positive links. To allow for this, we simply inflated the 0.00001 estimate of the probability of a link purely by chance to 0.000012. What might be the effect of varying meshblock size on the duplicate method?

We constructed a sensitivity analysis for record linkage conducted across blocks of two sizes: n=50 and n=150. For the same range of underlying values of p shown in Table 5 we then calculated the number of mortality records that would have zero, one, and two census record links, separately for the two sizes of meshblocks. When we assumed that the proportion of mortality records with a true link (t in the duplicate method equations on page 54) was 0.5, there was no substantive difference between the actual number of false positive links and those estimated by just using the average block size of 100. When we assumed t was 0.1, the duplicate method *overestimated* the number of false positive links by greater than 5% when p was greater than 0.001 – these latter conditions probability did not occur in this record linkage project. Therefore, we concluded that under the conditions experienced in this project, there was no substantive bias incurred by the duplicate method due to variation in the actual block sizes about the average.

Summary

The duplicate method seems reasonably robust within the range of values that will be encountered in the NZCMS. If the number of triplicate links becomes large (say greater than 5% of duplicate links) then it will be important to calculate the number of mortality records involved in duplicate links with census records, rather than just use the number of DA pairs.

2.5 The analysis of bias in the record linkage

2.5.1 Bias by cause of death

The majority of the analysis of bias was for all deaths combined. Variation in the linkage for specific causes of death were also investigated. Groupings of cause of death were based on that in Murray and Lopezs' 1990 Global Burden on Disease study,[31] with consideration of modifications proposed by Tobias (1998),[32] and is shown in Table 6. Many analyses were carried out at a higher level of aggregation based on similarities in record linkage characteristics between groups in Table 6. Those higher level groupings were: cancer, cardiovascular and respiratory disease, injury, infection, and other causes.

Table 6: ICD codes for groupings of cause specific deaths used in this research

Cause of death	ICD codes
Cancer	140-209
Ischaemic heart disease	410-414
Cardiovascular disease (other than IHD)	390-409, 415-459
Respiratory	470-478, 490-519
Infection	001-139, 320-323, 390-392, 460-466, 480-487, 590, 595, 614-616, 680-686, 711, 771
Unintentional injury	800-949
Suicide	950-959, 980-989
Interpersonal violence	960-979, 990-999
Other	remaining ICD codes

2.5.2 Stratified analyses of bias by demographic and socio-economic variables: all deaths

It is likely that the probability of a decedent's mortality record being successfully linked to their correct census record is associated with demographic and socio-economic variables. The objective of these analyses is therefore to *quantify* the bias in the record linkage by socio-economic factors. This bias becomes a *follow-up bias* in the cohort study, such that associations of death with socio-economic factors may be under or

overestimated due to bias in the record linkage itself. For example, if lower socio-economic status decedents were 20% less likely to be linked to a census record than high socio-economic status decedents, then any subsequent associations of socio-economic status with death in the cohort study would be underestimated by 20%.

The bias can be estimated by comparing the distribution of variables for linked and unlinked mortality records, without requiring access to census data. The demographic variables available for such analyses included:

- sex
- age (calculated from dates of birth and death)
- ethnic group
- region
- time from census night to death.

Ethnic group was derived from that on the NHI File. For the univariate and stratified analyses, ethnic group was classified as Maori, Pacific, all remaining specified ethnic groups (the 'Rest'), and not specified. For the regression analyses, the latter two categories were collapsed to a modified 'Rest' category.

The socio-economic variables available in the analysis of bias are:

- small area deprivation using the NZDep91 scale (derived from the meshblock, and available for about 89.4% of mortality records)
- employment status. Whether or not a usual occupation was recorded on the death registration form (BDM28) may provide a crude measure of employment status, although this is problematic:
  - the occupation entered on the BDM28 is the decedent's usual occupation in life - the decedent may still have suffered periods of unemployment if an occupation was recorded
  - if no occupation was recorded, it may not reflect unemployment, but one of the following: the decedent was a home-maker; the decedent did not actively seek employment; the decedent was retired, and the undertaker did not enter the

persons previous main occupation; the decedent had a usual occupation, but it simply was not entered

  - the interpretation of no occupation on the BDM28 varies by age and sex: a 22 year old with no recorded occupation could have been unemployed or a university student; a middle-age women with no recorded occupation could have been unemployed or a home-maker
- occupational class using the NZSEI scale. Whilst less problematic in its interpretation than whether or not an occupation was recorded, the interpretation of occupational class is also likely to vary by age and sex, particularly sex.

The NZDep91 and NZSEI are the most important measures, each being direct measures of socio-economic status. The NZDep91 is an index of deprivation for small areas. The index combines ten variables from the 1991 census which reflect seven dimensions of material and social deprivation. Each variable is a proportion of people in the area (eg proportion unemployed). The distribution of the weighted sum of these proportions is split into deciles to yield an integer score between 1 and 10 for each small area. The small areas are typically one or two meshblocks. The NZSEI is a socio-economic index of occupational class for New Zealand, again derived from 1991 census data.[30] The index was developed to supersede the dated Elley-Irving social class scale. It was developed by assigning each minor occupational group of the 1990 New Zealand Standard Classification of Occupations a score which maximised a latent factor mediating the effect of education on income, controlling for age. The NZSEI score is then transformed to a standard distribution ranging from 10 to 90, and can be categorised into six ‘occupational classes’ for comparability with the Elley Irving or other international scales. An alternative occupational class grouping than that proposed by Davis et al (1997) is used in this project (see Appendix 1, page 143).

The third socio-economic measure used was whether an occupation is recorded, a proxy for employment status, but should be treated with circumspection as it is probably a poor proxy.

Stratified analyses were conducted for the above variables (demographic and socio-economic), with linkage (yes, no) as the dependent variable. Statistical testing (eg confidence intervals) are not reported for the stratified analysis as, given the large sample size, most differences are statistically significant - the size of the difference is of greater importance. The stratified analyses also informed the selection of regression models in the next step.

2.5.3 Multiple regression analyses of bias by demographic and socio-economic variables: all deaths

A series of generalised linear models with a log link (referred to as log-linear hereafter) were conducted in SAS version 6.12. to assess the bias incurred in the record linkage by socio-economic factors. Statistical testing was used for the more complex regression analyses.

The regression model used for all final analyses was a log-linear risk model on the outcome of linkage to a census record (1=yes, 0=no):

$$R(x) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_n x_n)$$

where:

$R(x)$ is the average risk of being linked to census record given covariates x
$x_1, x_2, \ldots., x_n$ are the covariates or interaction products (eg sex, age group)
$\beta_1, \beta_2, \ldots., \beta_n$ 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 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 in the cohort analysis to allow for follow-up bias by socio-economic factors.

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 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 'risk' of being linked to a census record. This model was then used as the baseline to examine whether a 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 was 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 modeled (sex, age, ethnic group, time period between census and death, NZDep91 score, whether occupation recorded, and NZSEI occupational class) were either the exposure of interest, or covariates that had strong prior justification for inclusion.

For many of the later models investigating the marginal impact of socio-economic measures, an interaction term from previous models involving just demographic factors was treated as the main effect. This was appropriate given that all variables were nominal.

For whether an occupation was recorded analyses were restricted to decedents aged 15-74 years. For NZSEI occupational class, analyses were further restricted to 25-64 years. Given the different interpretation of occupation between sexes, these analyses were conducted separately by sex.

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 modeling. The logistic modeling assisted determining which interaction products to retain.

The above analyses could be conducted simply by stratification rather than regression modeling, deriving the 'actual' linkage rate by strata given that the mortality records analysed were the same target population of deaths for the cohort study. However, there were several limitations to using stratification, including:

- the SNZ protocol was that all absolute cell sizes must be random rounded to a multiple of three - the observed percentage linked in small cells will therefore be inaccurate
- the assigned ethnic group on mortality data is not equivalent to that on census data - regression modeling to 'smooth' out estimates by strata of ethnic group may be preferable to using actual results from health data ethnic group strata
- like ethnic group, the socio-economic measures available for all mortality records were not the same as those available for the census records - regression modeling may be more accurate at quantifying the follow-up bias by socio-economic factors (as an underlying dimension), rather than using the actual strata by socio-economic measure available on health data
- census records with either a census night dwelling of 'private hospital', 'public hospital', or 'rest-home', or simply a non-private census night dwelling, will be excluded from many of the cohort analyses. But we were unable to conduct an analysis of bias on a similar restricted set of mortality records as this variable was not recorded in the mortality file. Regression modeling to smooth out estimates by strata may, again, be preferable compared to using actual health data strata.

Chapter 3: Results – record linkage

The results are presented under the following headings:

- final output from the record linkage
- development of the record linkage strategy
- positive predictive value (PPV) estimates.

Subsequent sections present the analysis of bias in the record linkage. *The heading order in this section does not reflect the chronological order of the work: the **final** output from the record linkage was a result of the work detailed in the sections on the development and PPV, but is best presented first for clarity.*

3.1 Final output from the record linkage

3.1.1 Data flow of mortality and census records

42,229 mortality records were received from New Zealand Health Information Services, with inclusion criteria as described in the Methods Section (page 31). All but 46 of these mortality records were linked to a mortality record on the SNZ Vitals file. 17 NZHIS mortality records were linked to two SNZ Vitals file records, giving two separate mortality records for the same decedent. Despite non-New Zealand residents being excluded on the basis of NZHIS domicile codes, the SNZ Vitals file meshblock was coded as 'overseas usual residence' for a further 331 cases - they were excluded. One of the 331 overseas residents was for one of the 34 duplicate mortality records 'created' from the original 17 NZHIS mortality records. A decision was made to retain the 33 (34 minus 1) remaining duplicate mortality records, in case the true link could be established later. (This was not possible, and all 33 duplicate records were eventually discarded). Thus 41,915 mortality records were submitted to the record linkage (42,229 - 331 +17).

The record linkage involved submitting the mortality and census files to Automatch®, and deriving three output files:

- linked mortality and census records
- unlinked census records
- unlinked mortality records.

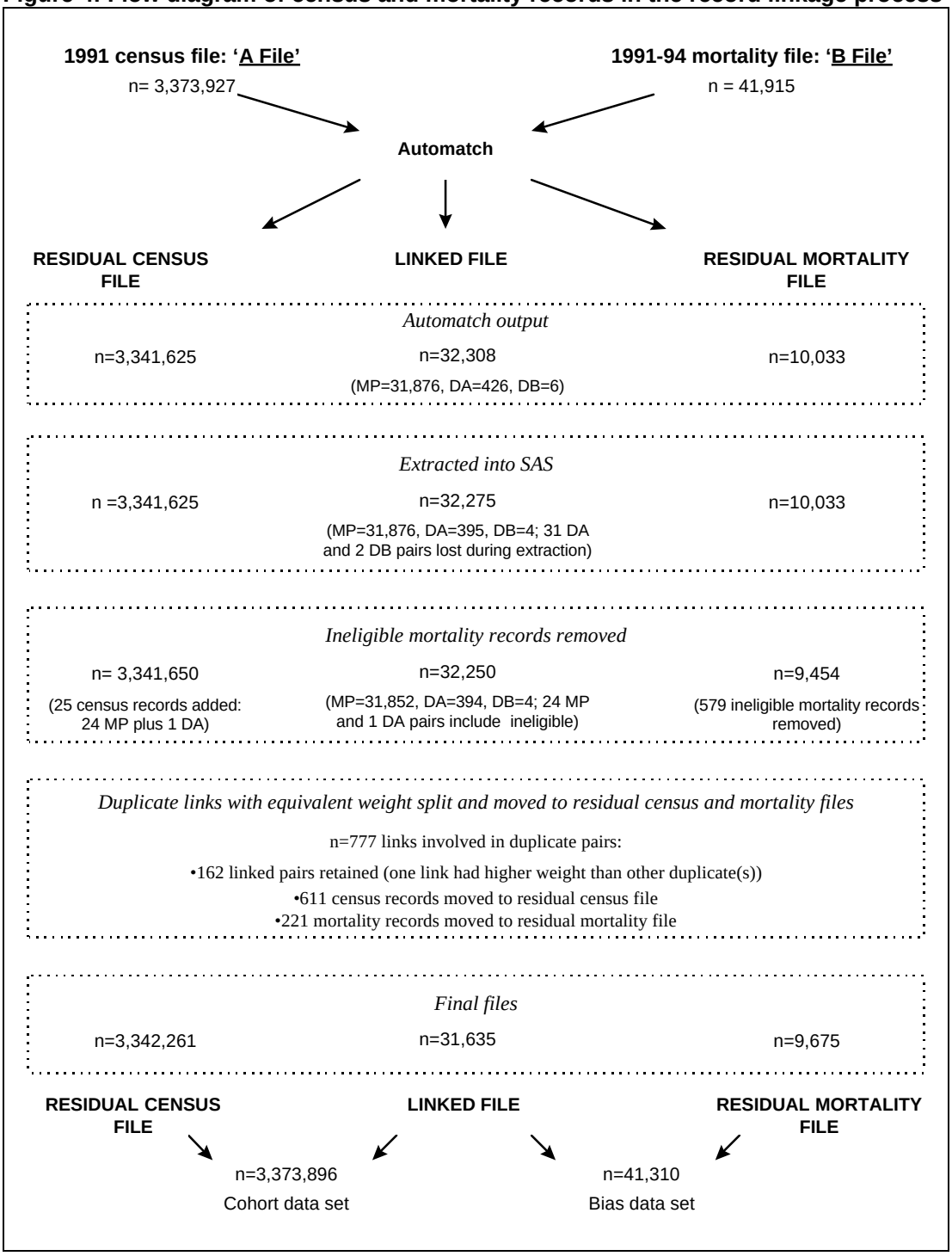
The flow of the mortality and census records is shown in Figure 4.

The total number of observations in the linked file of the Automatch® output (n=32,308) exceeds the number of either unique mortality records or unique census records due to the presence of links of one mortality record with two or more census records (DA pairs), and links of one census record with two or more for mortality records (DB pairs). Each DA and DB pair is associated with one MP pair, and the MP pair will be the highest weight pair. If the DA or DB pairs have the same weight as the MP pair, then both the MP and DA, or DB, pair have an equivalent probability of being a true link. If three census records link to one mortality record, then there would be one MP pair and two DA pairs all containing the same mortality record. Subtracting the number of DA pairs (n=426) from the total number of links (n=32,308), and adding the number observations in the residual mortality file (n=10,033) gives 41,915 - the total number of mortality records submitted. A similar calculation can be done to give the total number of census records submitted, but instead subtracting the number of DB pairs (n=6).

During the extraction of data from Automatch® to SAS, 31 DA pairs and 2 DB pairs were ‘dropped’. The reason was not determined, and it was not detected until much of the processing of the links had been conducted in SAS. Given the large amount of time and resource that would have been required to re-run the final match-run strategy, and the lack of certainty that the same problem would not recur in any further extraction, these 33 observations were accepted as lost. The overall impact was minor, being 2 out

of 41,915 submitted mortality records (0.005%) and 31 of 3,373,927 submitted census records (0.0009%).

Figure 4: Flow diagram of census and mortality records in the record linkage process



MP = Match pair of one census and one mortality record
DA = Duplicate A pair, that is a pair including a census record not involved in a MP pair, but including a mortality record involved in a MP pair. Any MP pair and DA pair(s) involving the same mortality record are kept 'associated' by Automatch® for latter clerical review.
DB = Duplicate B pair, the reverse of a DA pair.

The mortality data-request to NZHIS was for people aged 0-74 on census night. However, the data actually included people born up to a year after the census (n=532) - this was detected during the final match-run. Also included (knowingly) in the submitted mortality records were the 33 observations for the 17 NZHIS mortality records with two SNZ Vitals file links, and 38 decedents who actually died on 5 March 1991 (census day). Inspection of records suggested there would be little chance of successfully teasing apart the 17 duplicates. Further investigation also suggested that the likelihood of someone dying on census day having had a census completed by them (or on their behalf) was remote. Therefore, these 603 (532 + 33 + 38) 'ineligible' mortality records were removed from the data. Calculations (not presented here) suggested that inclusion of these 603 ineligible records had no effect on the probability of a true link being found for the remaining eligible mortality records. Therefore, there was no justification to repeat the final match-run of the record linkage.

Automatch® does not allow the DA or DB pair(s) *and* the associated MP pair to all be discarded together - instead one link has to be accepted as the correct MP pair. For DA and DB pairs with an equivalent weight to the associated MP pair, the DA and DB pair(s) *and* the MP pair were to be discarded. This operation had to be conducted in SAS. 777 linked pairs were involved in a MP/DA or MP/DB association. 162 MP pairs had a higher match weight than the associated DA or DB pair(s), and were therefore retained as the 'best link'. The remaining 615 links were separated into 611 unique census records, and 221 unique mortality records.

The final size of the linked file was 31,635, and included links from all eight passes of the final match-run strategy. The sum of the linked file and residual census file records was 3,373,896, 31 less than the original census file size due to the loss of 31 DA pairs during extraction from Automatch®. The sum of the linked file and residual mortality file records was 41,310 - two less than the number of eligible mortality records due to the loss of 2 DB pairs during extraction from Automatch®.

Finally, a further data management issue requires stating for completeness. The census file counts in Figure 4 are for New Zealand residents only. There were actually 162,189 further census records submitted to Automatch®: 101,166 absentee census

records and 61,023 overseas residents. However, none of these census records would have been available for linkage to a mortality record as they did not have 'legitimate' usual residence meshblock or CAU codes. The only effect on the record linkage would have been to cause a slight underestimate (about 3%) for all of the *u* probabilities, as 3% of the submitted census records (101,166 absentee records) had no value for any of the matching variables. This mild underestimate of the *u* probabilities would have slightly widened the distribution of total weight scores for all possible comparisons (i.e. a slight increase in distance between the two peaks), but it would not have changed the ranking of comparisons by weight, and thus would not have altered the links accepted as true links.

3.1.2 Final match-run strategy

The final match-run strategy, and number of links by pass, is presented in Table 7. The results are following the use of MPROB on the full match-run, to allow for value specific *m* probabilities. The majority of the linked mortality records were identified on the first pass (25,311, or 61.27% of the total 41,312 eligible mortality records). For all eight passes, 76.6% of mortality records were linked to a census record. The details of each pass, and the developmental work to determine the best configuration and order of the passes, is presented in a subsequent section (heading number 3.2.2, page 83). Brief details of each pass are in the footnotes to Table 7.

Table 7: Final match-run strategy, using post-MPROB

Pass and blocking variable(s)	Main match specifications	Matching variables	Links (% of eligible mortality records)†	
1. Meshblock	• Match cut-off weight 23.0 • +/- 1 tolerance for dd, mm, and yyyy	• Sex, dd, mm, yyyy, and ethnic group from both NMDS and NHI • Birthplace from NMDS	25,311	(61.27%)
2. Vitals-CAU, and month of birth	• Match cut-off weight 23.0 • +/- 1 tolerance for dd and yyyy	• Sex, dd, yyyy, and ethnic group from both NMDS/NHI • Birthplace from NMDS	3473	(8.41%)
3. Post-CAU, and month of birth	(As for pass 2)	(As for pass 2)	1117	(2.70%)
4. Pre-CAU, and month of birth	(As for pass 2)	(As for pass 2)	340	(0.82%)
5. NHI-CAU, and month of birth	(As for pass 2)	(As for pass 2)	416	(1.01%)
6. Meshblock	• Clerical review weight range 20.0-22.9 • +/- 1 tolerance for dd, mm, and yyyy	(As for pass 1)	429	(1.04%)
7. Meshblock	• Clerical review weight range < 20.0 • no tolerance for dd, mm, and yyyy	(As for pass 1)	91	(0.22%)
8. Vitals-CAU and month of birth	• Clerical review weight range < 23.0 • **no** tolerance for dd, mm, and yyyy	(As for pass 2)	458	(1.11%)
Total			**31,635**	**(76.58%)**

The source of the blocking variable for census records on all passes is that derived from the census usual residence. The source of the blocking variable for mortality records varies: meshblock = the meshblock from the SNZ Vitals file; Vitals-CAU = the CAU from the SNZ Vitals file; post-CAU = the CAU from the NMDS file for the health event (if any) immediately after the 1991 census; pre-CAU = the CAU from the NMDS file for the health event (if any) immediately before the 1991 census; NHI-CAU = the CAU from the NHI file.

† Links here are those remaining after full data cleaning as depicted in Figure 4.

dd = day of birth; mm = month of birth; yyyy = year of birth.

3.1.3 Accuracy of the record linkage: false positives and false negatives

The estimated positive predictive value and number of false positives for the first five passes are shown in Table 8. Two methods were used to estimate the positive predictive value: the chance method and the duplicate method. (The methods were described in the Methods section, and more detailed results for the two methods are included in following sections.)

The overall PPV for the first five passes was estimated to be 97.8% by the chance method, and 98.1% by the duplicate method. The close agreement between the chance and duplicate method allows confidence in the robustness and accuracy of both methods. It was not possible to estimate the PPV directly for the last three clerical review passes, but it was probably in the range of 80% to 90% based on work undertaken in the development of the clerical review rules (heading 3.2.3, page 85). Assuming it was 85% for these three final passes, then the PPV for all eight passes combined was about 97.3% to 97.7%.

For practical purposes of comparison, the eight passes can be divided into three groups:

- very high PPV (greater than 99.5%; pass 1; 80.0% of all linked mortality records)
- high PPV (approximately 90%; pass 2-5; 16.9% of all linked mortality records)
- moderate PPV (80-90%; passes 6-8; 3.1% of all linked mortality records).

Table 8: Positive predictive value (PPV) and expected number of false positives (E[FP]) for passes 1 to 5 of the final match-run, using both the duplicate method and chance method

Pass	Link pairs	Chance method E[FP]	Chance method PPV	Duplicate method E[FP]	Duplicate method PPV
1 Meshblock, wt>30.0	23000	22	99.9%	48	99.8%
Meshblock, wt<30.0	2311			37	98.4%
2 Vitals-CAU	3473	365	89.5%	274	92.1%
3 post-CAU	1117	130	88.4%	134	88.0%
4 pre-CAU	340	52	84.9%	39	88.5%
5 NHI-CAU	416	81	80.5%	41	90.1%
Totals †	**30657**	**687** †	**97.8%**	**573**	**98.1%**

† For the chance method, the totals include the 37 estimated false positives by the duplicate method below the exact cut-off (30.0) for pass 1 to allow comparability.

The number of false negative links are approximated, although mildly overestimated, by the 9,677 mortality records not linked to a census record (23.4% of all mortality records). This will be an overestimate of the true number of false negatives as:

- some decedents would not have been in New Zealand on 1991 census night

- some decedents would not have completed the census, despite being in New Zealand on 1991 census night.

The 9,677 unlinked mortality records also includes 221 mortality records that were linked to a census record, but were rejected as there was a duplicate link with the same weight meaning it was impossible to select the most likely link (ie. there was only a 50:50 chance of selecting the true link, so they were both discarded).

Taking the above into account, it seemed reasonable to conclude that:

- about 20% of the mortality records were false negative links (i.e. they were not linked when in fact there was a true link somewhere in the census file)
- about 2.5% of the linked mortality records were false positive links
- and, therefore, about 22.5% of mortality records were either incorrectly linked or incorrectly not linked.

3.1.4 Final *u* and *m* probabilities

Final *u* and *m* probabilities for the full match-run are shown in Table 9. The *m* probabilities are those determined by MPROB.

Table 9: *u* and *m* probabilities, and agreement and disagreement weights for matching variables for the final match-run (m probabilities determined by MPROB)

Matching variable		m probability	u probability	Agreement weight	Disagreement weight
Sex (NMDS)	Male	1.00	0.48	1.05	-8.83
	Female	1.00	0.49	1.02	-8.56
Day of Birth †	Range	0.96 to 0.99	0.02 to 0.03	4.90 to 5.72	-6.39 to -4.45
Month of Birth †	Range	0.98 to 0.99	0.07 to 0.09	3.50 to 3.72	-7.28 to -5.88
Year of Birth	1900	0.99	0.00	10.55	-6.14
(NMDS,	1910	0.99	0.00	8.24	-6.13
examples by	1920	0.99	0.01	7.03	-6.24
decade)	1930	0.99	0.01	6.83	-6.52
	1940	0.99	0.01	6.56	-6.18
	1950	0.99	0.01	6.17	-7.06
	1960	0.99	0.02	5.91	-6.09
	1970	0.99	0.02	5.91	-7.46
	1980	0.98	0.01	6.09	-5.59
	1990	0.99	0.02	5.92	-7.51
Ethnic group	Maori	0.81	0.12	2.72	-2.22
(NHI)	Pacific	0.83	0.04	4.25	-2.53
	Asian	0.58	0.03	4.31	-1.20
	Other	0.00	0.00	0.01	0.00
	European	0.89	0.76	0.22	-1.10
Ethnic group	Maori	0.73	0.12	2.56	-1.69
(NMDS)	Pacific	0.65	0.04	3.90	-1.45
	Rest	0.96	0.80	0.27	-2.53
Birthplace	NZ	0.99	0.80	0.31	-4.40
(NMDS)	Australia	0.96	0.01	6.12	-4.62
	British Isles	0.97	0.07	3.83	-4.97
	Europe	0.97	0.01	6.13	-4.83
	Pacific Is	0.96	0.03	5.10	-4.55
	Africa	0.94	0.00	8.41	-4.02
	America	0.93	0.01	7.50	-3.82
	Asia	0.92	0.02	5.72	-3.65
	Other	0.88	0.00	9.93	-3.03

† Both NMDS and NHI.

3.2 Development of the record linkage strategy

Section 3.1 above summarises the output from the record linkage. This section outlines in greater detail how the record linkage strategy was developed.

3.2.1 Determining the match cut-off

The match cut-offs were decided by a combination of looking at histograms of the distribution of link pairs by weight, and inspection of link pairs.

3.2.1.1 Meshblock

A trial record linkage was conducted for the meshblock pass only, using the same match specifications as for pass 1 of the final match-run, that is:

- matching variables of sex, dd (day of birth), mm (month of birth), yyyy (year of birth) and ethnic group from both the NMDS and NHI files
- NMDS and NHI file matching variables for the same underlying characteristic (eg sex) considered separately, not as an array
- matching variable of birthplace from the NMDS file only
- tolerance of +/- 1 for dd, mm, and yyyy (eg dd of 14 on one file and 15 on the other file would gain a partial agreement weight, that is a weight score half way between the agreement and disagreement weight).

MPROB was conducted for this trial linkage to generate value specific *m* probabilities, using a weight cut-off (to determine linkage status) of 25.0.

The distribution of links by weight is shown in Figure 5. The number of links for weights of less than 8.0 are not shown, but are vastly more numerous than those with a weight above 8.0. Thus, the peak at 30 to 40 represents the smaller peak of the bi-modal distribution of links, that is those links with a high probability of being true links. The trough from 15 to 30 represents the ‘grey zone’ of uncertain links. Above a weight of 30.0, most links were exact links (ie. each matching variable agreed exactly). If they were not an exact match, then it was usually due a disagreement on the ethnic group or country of birth variable.

Figure 5: Histogram of the number of links by Automatch® weight for a trial record linkage of the meshblock pass

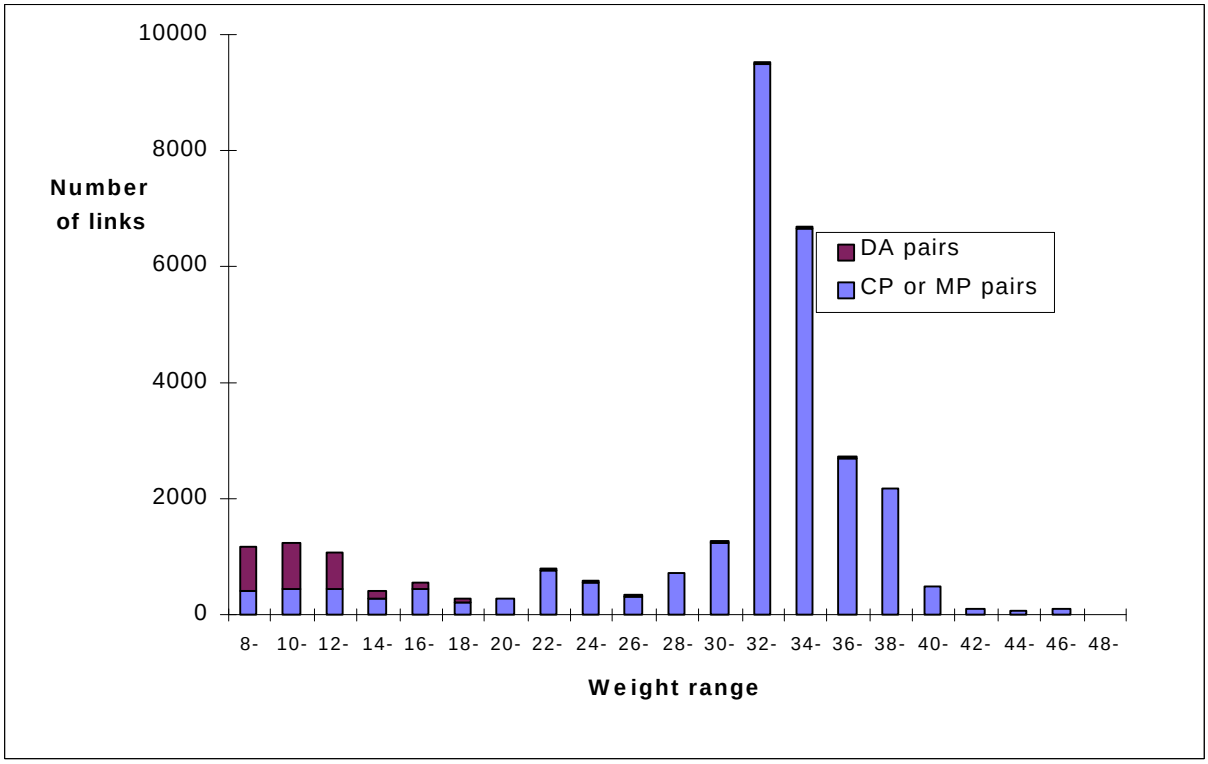
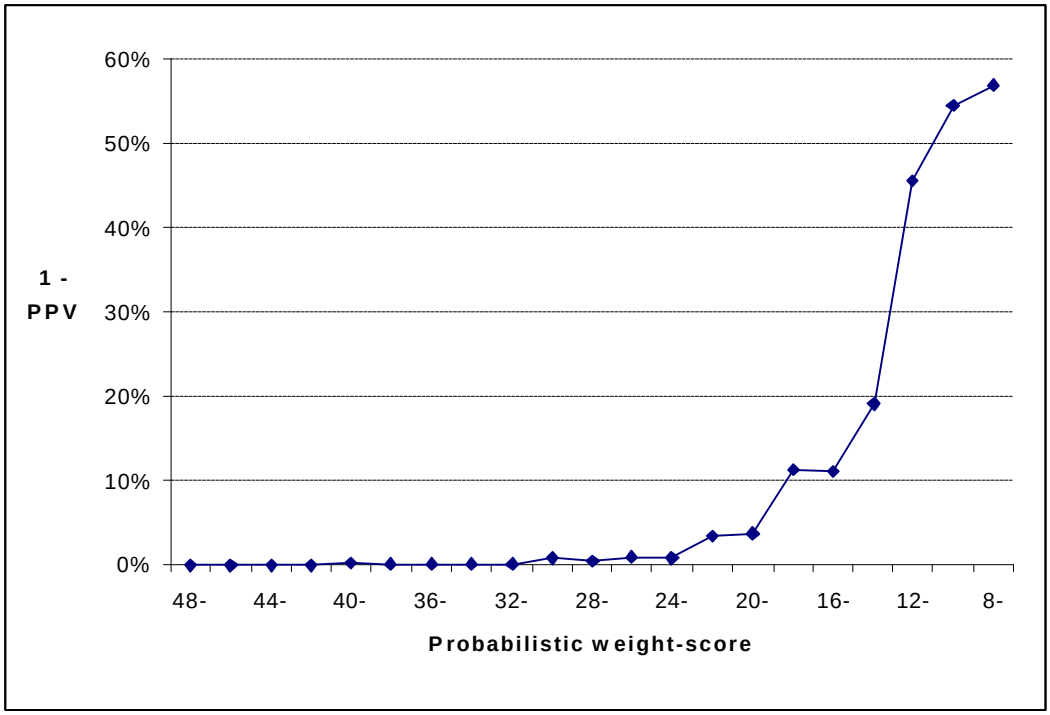


Figure 6 shows a plot of the percentage of linked records estimated to be false positives by the duplicate method, against the Automatch® weight (grouped). For each 2-unit range of probabilistic weight score (eg 34.0-35.99) above a weight of 30.0, 1-PPV is consistently below 1%. From 30.0 to 20.0, that increases to about 5%.

Figure 6: Percentage of links that are false positives (1-PPV) for an initial meshblock pass, estimated by the duplicate method.



Against a backdrop of the histogram in Figure 5, and the percentage of false positives by weight in Figure 6, a random sample of links between a weight of 8.0 and 25.0 were visually inspected. It was determined that above a weight of 23.0 no links would be rejected by clerical review. That is, above a weight of 23.0 a human could add nothing to the probabilistic record linkage, and all links would be accepted.

The match cut-off for the meshblock pass was therefore set at 23.0.

3.2.1.2 CAU

The CAU passes were also blocked by month of birth. That is, each CAU block was further broken down into 12 blocks of month of birth, using the census and NMDS file month of birth. This blocking by both CAU and month of birth was to prevent block overflow (too many comparisons within anyone block).

The matching variables at the CAU-level were:

- sex, dd, yyyy and ethnic group from both the NMDS and NHI files

- birthplace from the NMDS file only.

Unlike the meshblock pass, no +/-1 tolerance was allowed around day and year of birth.

The total weight assigned by Automatch® for CAU passes was equivalent to that assigned at the meshblock level, minus twice the weight assigned for an agreement on month of birth (2×3.5 = 7.0; month was included twice in the meshblock pass - once for the NMDS file and once for the NHI file). If exact (or near-exact matches) are defined in the same way as at the meshblock pass, the threshold moves from 30.0 to 23.0. (Note that the actual threshold used in the meshblock pass was 23.0, but 'near exact' matches were only above a weight of 30.0.) 'Near-exact' matches is an appropriate concept here, as some types of disagreement for the same individual are not just a function of coding error, but recognised systematic biases in the way data is collected in New Zealand. For example, it is common for one individual to self-report as Maori on the census, yet be coded as non-Maori on mortality data. The links above 30.0 at the meshblock-level, and hence above 23.0 at the CAU-level, that are not exact matches for all matching variables tend to be of this type. To set a match cut-off above 23.0 at the CAU-level, whilst improving the PPV, would also bias against certain types of decedents being linked, including:

- individuals with common values for matching variables (eg. male, young, New Zealand European, born in New Zealand), but with one minor error (eg census day of birth disagreeing with just one of either the NMDS or NHI day of birth) would not be linked
- individuals coded as Maori on the census, but coded as non-Maori (the 'Rest') on mortality data, would not be linked when they were very likely the same person.

On theoretical grounds, 23.0 was therefore the maximum match cut-off that could be used at the CAU-level without introducing likely bias. On the other hand, lowering the match cut-off would probably result in an unacceptable PPV.

The match cut-off for CAU-level passes was set at 23.0.

3.2.2 Determining the pass order

The ordering of the CAU passes that was thought to be the most theoretically robust a priori was:

- post-CAU: CAU for the health event (if any) immediately after the census
- pre-CAU: CAU for the health event (if any) immediately before the census
- NHI-CAU: CAU for the NHI file
- Vitals-CAU: CAU from aggregating the meshblock obtained from the SNZ Vitals file.

The main reason to use the CAU passes was to try and link decedents that had moved between census night and death, and would therefore have a SNZ Vitals file assigned meshblock different to their census usual residence meshblock. Hence, the post-CAU and pre-CAU codes from the NZHIS health data were thought to be the two high priority CAU passes. The NHI-CAU followed, as in some instances it may be a residence preceding the residence at death. The Vitals-CAU was thought to be the lowest priority, being the aggregate of the meshblock on the SNZ Vitals file for the usual residence at death, or the directly coded CAU if the coders were unable to assign a meshblock in the first instance. Mortality records were only submitted to the post-CAU, pre-CAU, and NHI-CAU passes if the CAU code was different from their Vitals-CAU code.

MPROB was not conducted for the trial match-runs to test the ordering of CAU passes due to time constraints.

The results for the initial trial match-run strategy found a higher number of MP pair links for the Vitals-CAU than the three other passes. Additionally, the estimated PPV for the Vitals-CAU pass (90%) exceeded those estimated for all three other CAU passes (78-86%). The reasons for this unexpected good performance of the Vitals-CAU pass were probably that:

- 10% of all mortality records (n=4,373 of the original 41,915 mortality records) only had a CAU (no meshblock) assigned following linkage with the SNZ Vitals file.

Thus, these records could not be linked on the first meshblock pass, but were able to be linked on the Vitals-CAU pass

- the degree of misclassification of meshblock codes, but misclassified to another meshblock in the same CAU, may have been greater than expected. Indeed, discussions with the SNZ coders responsible for assigning the meshblock/CAU codes for the SNZ Vitals file revealed that when more than one meshblock was available for selection, the 'most likely' meshblock was coded rather than just assigning the CAU code.

A second trial match-run therefore placed the Vitals-CAU pass as the first CAU pass. Results are shown in Table 10. All mortality records remaining unlinked from the first meshblock pass were submitted to the Vitals-CAU pass, but only mortality records with a CAU code different to those in the preceding pass(es) were submitted to the post-CAU, pre-CAU, and NHI-CAU passes. (The methodological details of the chance and duplicate method for estimating the number of false positives are presented in the Methods Section. Full details of calculations are presented for the final PPV estimates in a following section.)

Table 10: Second trial match-run of CAU pass order: number of MP pairs, and estimated number of false positive MP pairs

	Mortality records	Average block		Exact method		Duplicate method		
Pass	submitted*	size	MP pairs	E[FP]	PPV	DA pairs	E[FP]	PPV
Meshblock	37479	132	25195	70	99.7%	175	76	99.7%
Vitals-CAU	16716	204	3766	381	89.9%	127	371	90.1%
post-CAU	5385	212	1199	127	89.4%	55	170	85.8%
pre-CAU	1945	217	375	49	86.9%	13	47	87.5%
NHI-CAU	2845	217	438	75	82.8%	13	63	85.6%
Totals			30973	717	97.7%	383	651	97.9%

E[FP] = estimated number of false positive MP pairs. For the chance method at the meshblock pass, this includes the number of false positives estimated by the duplicate method for the weight range 23.0 to 30.0.
* Mortality records submitted to each pass were estimated by multiplying the Automatch® stated outputs of [average block size] and [average number of mortality records per block processed].

The PPV of the Vitals-CAU pass in the second trial run was maintained, and the PPV for the three remaining passes *improved*. The reason for this improvement was that by

placing the Vitals-CAU pass first, there were fewer residual mortality records submitted to subsequent passes that could give rise to a false positive link.

It was decided that no improvement could be made to the order of the first five passes shown in Table 10. Additional passes for clerical review of the meshblock and Vitals-CAU pass were considered appropriate, and their development is outlined in the next section.

3.2.3 Determining the clerical review rules for passes 6-8

There were likely to be many true links below the match cut-offs in the first five passes: the task was to find them by clerical review without incurring many false positives. General inspection of links below the match cut-off for the first meshblock pass suggested a number of possible types of true link:

- an obvious transposition (eg. day and month of birth swapped)
- disagreements for ethnic group or country of birth
- links where one of the NMDS or NHI sex, dd, mm, and yyyy variables disagrees with the census variable, but not both (eg NHI dd disagrees with census dd, but NMDS dd agrees)
- links where dd, mm, or yyyy are the same for the NMDS and NHI variable, but disagree by +/- 1 with the census variable
- links where dd, mm, or yyyy are the same for the NMDS and NHI variable, but disagree by more than +/- 1 with the census variable.

A random sample of links (both MP and DA links) in the weight range 8.0 to 24.9 for the first meshblock pass were further inspected to investigate the likely accuracy for the above different types of links. This random sample was made up of 14 DA pairs and 9 CP pairs for each weight group (eg. 16.0-16.9). (A CP pair is the same as a MP pair, except is in the clerical review range rather than above the match cut-off). Using the duplicate method, very crude estimates of the likely PPV could be made for different clerical review rules by back-calculation from the number of DA and CP pairs fitting the rule.

These back-calculations suggested that links where one of sex, dd, mm, or yyyy disagreed between the NHI and NMDS file, and hence only one of either the NHI or NMDS file variables agreed with the equivalent census variable, had a high PPV - perhaps about 99%. When the NMDS and NHI variables agreed, a disagreement of +/- 1 for one of the dd, mm, or yyyy variables with the equivalent census variable had a fair PPV - perhaps 85-90% when ethnic group and country of birth agreed exactly, less if there was minor disagreement of ethnic group or country of birth. There were not enough transpositions to make any quantitative estimation of accuracy, but subjectively they seemed to be highly likely to be true links.

It was also noted that some of the NHI and NMDS files seemed to be incorrectly linked for the decedent, let alone any links with census records. As the final analyses in this study were to be between census measured exposures and NMDS Death Event File outcomes, it was more important that the NMDS variables agreed with the census variables.

On the basis of these crude back calculations, and deliberation between the principal investigator and SNZ staff, clerical review rules were developed as shown in Table 11.

Table 11: Final clerical review rules for record linkage

No.	Clerical review rule	Pass[‡]
1.	Accept CP pairs with an <u>obvious</u> transposition. For example: (111) (21 21 **9)** (9 9 **21)** (1925 1925 1925) (5533) (11) where: • the clusters are the (sex), (dd), (mm), (yyyy), (ethnic group), and (country of birth) variables respectively • for sex, dd, mm, and yyyy clusters, the order of variables is census, NHI, NMDS • for ethnic group, the order of variables is census, NHI, census, NMDS, • for country of birth, the order of variables is census, NMDS.	6,7,8
2.	**Accept CP pairs where either (but not both) the NHI or NMDS sex variable disagrees with the census sex variable, and dd, mm, and yyyy were exact matches, and ethnic group and country of birth were acceptable.** For example: (1**2**1) (21 21 21) (9 9 9) (1925 1925 1925) (5533) (11) where acceptable disagreement is: • for ethnic group: 5_33; 1513; 1113; 1511; 5433; 5133; 2223; and 1411[†] • for country of birth: 11; 12; 13; 21; 31.	6,7
3.	**Accept CP pairs where <u>either</u> (but not both) the NHI or NMDS dd, mm, or yyyy variable disagrees with the equivalent census variable. This disagreement can be for one, two, or all three of the dd, mm, and yyyy clusters, <u>so long as no more than one disagreement was between a NMDS and census variable</u>. There must be at least acceptable agreement on ethnic group and country of birth.** For example: (111) (21 21 **9**) (9 **13** 9) (1925 1925 1925) (5533) (11) (111) (21 21 **9**) (9 **13** 9) (1925 1925 1925) (**1513)** (11) (111) (21 **24** 21) (9 **13** 9) (1925 **1938** 1925) (1111) (11) But not: (111) (21 21 **24**) (9 9 **13**) (1925 **1938** 1925) (1111) (11)	6,7,8
4.	**Accept CP pairs where <u>both</u> the NHI and NMDS dd, mm, or yyyy variable disagrees by +/- 1 with only one equivalent census variable, and there is <u>exact</u> agreement on sex, ethnic group and country of birth, and the CP pair weight is greater than 20.0 on the meshblock pass.** For example: (111) (**20** 21 21) (9 9 9) (1925 1925 1925) (3355) (11) but not: (111) (**20** 21 21) (9 9 9) (1925 1925 1925) (**1315**) (11)	6

[†] For the first two digits of each four digit ethnic group cluster (the 'NHI' cluster), 1=Maori, 2=Pacific, 3=Rest (including European). For the last two digits (the 'NMDS' cluster), 1=Maori, 2=Pacific, 3=Asian, 4=Other (but the majority are 'Other European, NZHIS code 54), 5=European.

[‡] Pass 6= meshblock blocking, clerical review range 20.0-22.99, tolerance +/- 1 on dd, mm, yyyy. Pass 7= meshblock blocking, clerical review range 12.0-19.99, no tolerances. Pass 8= Vitals-CAU blocking, clerical review range 8.0-22.99, no tolerances.

In order to meet these clerical review requirements for the meshblock pass, it was sensible to run the meshblock pass twice. The first run would retain the tolerance of +/- 1 for the dd, mm, and yyyy variables to allow clerical rule 4 to be applied to

mortality records with a weight greater than 20.0. The second run would drop the tolerances to prevent unnecessary clerical review of links with a +/- 1 disagreement at the expense of other types of links.

Clerical review of the Vitals-CAU pass was considered appropriate under strict conditions, that is only mortality records fitting clerical review rules 1 and 3. Clerical review of the other CAU passes was not conducted in the belief that it would have been more inaccurate, and had a relatively small marginal gain.

Table 12: Match specifications and clerical review rules for passes 6 to 8

Pass	Match Specifications	Matching Variables	Clerical review rules
6. Meshblock	• Clerical review range 20.0-22.9 • +/- 1 tolerance for dd, mm, and yyyy	• Sex, dd, mm, yyyy, and ethnic group from both NMDS/NHI • Birthplace from NMDS	1, 2, 3, 4
7. Meshblock	• Clerical review range < 20.0 • **no** tolerance for dd, mm, and yyyy	(As for pass 6)	1, 2, 3
8. Vitals-CAU and month of birth	• Clerical review range < 23.0 • **no** tolerance for dd, mm, and yyyy	• Sex, dd, yyyy, and ethnic group from both NMDS/NHI. • Birthplace from NMDS	1, 3

Note that for pass 8, some links were exact agreements on sex, dd, mm, and yyyy, but with minor disagreement on ethnic group and country of birth. They were also accepted.

3.3 Workings for the positive predictive value estimates

3.3.1 PPV estimates by pass: chance and duplicate methods

Summary results for the PPV estimated by the exact and duplicate method by pass were presented previously in Table 8 (page 76). The full information necessary to calculate these estimates is presented in Table 13 and Table 14. The chance method

calculations were re-iterated to allow for an improved estimate of the number of true links by pass, but only made a small amount of difference.

Table 13: Workings for final PPV estimates by pass: chance method

Pass	Mortality records submitted (MRS)	Match Pairs (MP)	Average block size (AB)	Probability mortality record linking purely by chance (Pr)	Iteration 1		Iteration 2		Iteration 3	
					Estimated number FP ($E[FP]_1$)	PPV_1	$E[FP]_2$	PPV_2	$E[FP]_3$	PPV_3
				$1.2\times10^{-5} \times AB$;‡ $1.2\times10^{-5} \times 12 \times AB$#	(MRS - MP) × Pr	1 - ($E[FP]_1$ / MP)	(MRS - MP + $E[FP]_1$) × Pr	1 - ($E[FP]_2$ / MP)	(MRS - MP + $E[FP]_2$) × Pr	1 - ($E[FP]_3$ / MP)
Meshblock†	36940	23068	132	0.0016	22	99.9%	22	99.9%	22	99.9%
Vitals-CAU	15862	3552	204	0.0294	362	89.8%	373	89.5%	373	89.5%
post-CAU	5441	1157	212	0.0305	130	88.7%	134	88.4%	135	88.4%
pre-CAU	1992	350	217	0.0313	51	85.3%	53	84.9%	53	84.9%
NHI-CAU	2966	421	217	0.0313	80	81.1%	82	80.5%	82	80.5%
Totals #		**30865**			**699**	**97.7%**	**718**	**97.7%**	**719**	**97.7%**

† Only for those links with a weight above 30.0.
‡ Formula for meshblock where 1.2×10^{-5} is the probability of an exact link between any one mortality record and any one census record, purely by chance (see Methods section).
* Formula for CAU. As for meshblock, except multiplied by 12 to allow for the loss of month of birth as a matching variable.
# Include numbers determined for the weight range 23.0-29.9 on the meshblock pass by the duplicate method to allow comparability with Table 14.

Table 14: Workings for final PPV estimates by pass: duplicate method

Pass	Automatch® output			Estimates required for duplicate method			Duplicate method results	
	Mortality records submitted (MRS)	Match Pairs (MP)	DA Pairs (DA)	Estimated no. mort. records with single census link Mesh = MP – DA CAU = MP – (0.9 × DA) †	Estimated no. mort. records with duplicate census link Mesh = DA CAU = 0.8 × DA †	Estimated no. mort. records with no census link Mesh = MRS - MP CAU = MRS - MP †	Estimated number false positive links (E[FP]) (SeeSection 2.4.3, page 53, for formulae)	Estimated PPV 1 – (E[FP] / MP)
Meshblock, wt >=30.0	36940	23068	78	22990	78	13872	48	99.79%
Meshblock, wt < 30 ‡	na	2317	103	2214	103	11553	37	98.41%
Vitals-CAU	15862	3552	95	3467	76	12310	274	92.28%
post-CAU	5441	1157	42	1119	34	4284	134	88.41%
pre-CAU	1992	350	10	341	8	1642	39	88.89%
NHI-CAU	2966	421	8	414	6	2545	41	90.28%
Totals		**30865**	**336**	**30545**	**305**		**573**	**98.14%**

† More than one DA pair per MP pair (i.e. three census links per mortality record) was common at the CAU-level. The actual number of MP pairs associated with DA pairs was about 0.9 times the number of DA pairs. Assuming that no mortality records were linked with four or more census records, then the number of mortality records linked with two census records (duplicates) was 0.8 × the number of DA pairs. Correspondingly, the number of mortality records linked with three census records (triplicates) was 0.1 × the number of DA pairs.

‡ To determine the number of false positive links on the meshblock pass *beneath* a weight of 30 requires calculating the first row of this table, then the number of false positives for the full meshblock pass combined (not shown), and subtracting the former from the latter.

The small number of observed DA pairs in the latter CAU passes makes the PPV estimates by the duplicate method unstable. Conversely, the chance method estimates remain precise and provided a useful relative ranking.

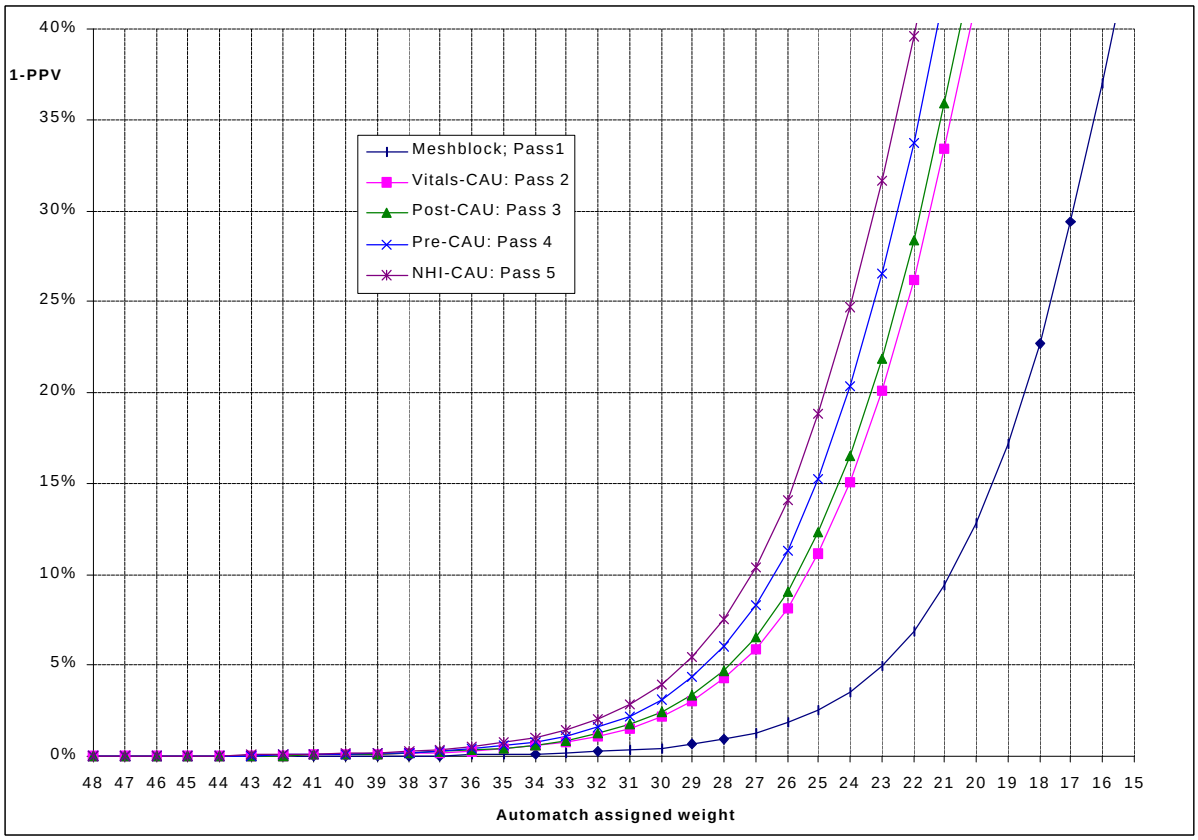
The PPV of the clerical review passes was not able to be estimated directly. However, it was probably about 80-90% for each of the three passes based on work undertaken during the development of the clerical review rules. Assuming it was 85%, the overall PPV for all 31,635 links on the eight passes was estimated to be 97.3% to 97.7%.

3.3.2 PPV estimates by pass: absolute weight method

The PPV by weight was estimated using the absolute weight method for each of the first five passes as shown in Figure 7. The absolute weight method assumes that DA and DB pairs are false positives. Conversely, the duplicate and chance methods developed for this project ignore the DA and DB pairs as they will not be accepted as links, as each mortality and census record can only be linked once. The result is that the absolute weight method will overestimate the number of false positives compared to that experienced in this research, particularly for the CAU passes where the relative number of DA and DB pairs is greater. Additionally, the absolute weight method is prone to bias as discussed in the Methods section (page 45).

These above caveats acknowledged, the absolute weight method will be precise, even if not highly accurate. As the likely biases above will be the same for each of the four CAU passes, the relative position of the curves in Figure 7 is likely to be a good measure of the *relative* accuracy of each of the four CAU passes. Therefore, the PPV at any given weight for the Vitals-CAU and Post-CAU passes as shown in Figure 7 are similar, the PPV for the NHI-CAU pass lowest, and the PPV of the pre-CAU pass intermediary. This corroborates the chance method that also predicted the same relative ranking of the four CAU passes.

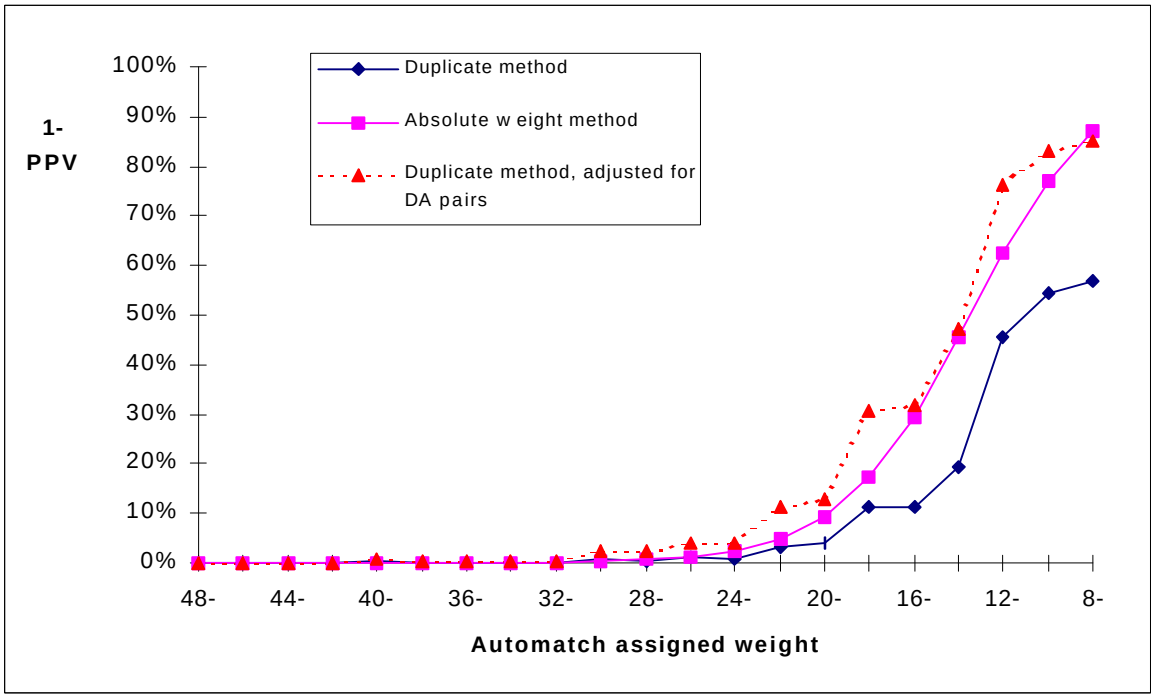
Figure 7: Percentage of links likely to be false positives (1-PPV) by weight for passes 1 to 5 by Automatch® assigned weight, using the absolute weight method



3.3.3 PPV estimates by weight for the meshblock pass (pass 1): duplicate method and validation by absolute weight method

Data to determine the PPV by the duplicate method below the match cut-off of 23.0 in the *final* match-run was not available, as Automatch® does not print out results below the nominated match cut-off. However, the necessary data was available for an earlier trial run and results are shown in Figure 8. Also included in Figure 8 are curves for the absolute weight method, and the duplicate method *adjusted* to include the observed number of DA and DB pairs as false positives. The latter two curves both include duplicate links as false positives, whereas the duplicate method only estimates the number of non-duplicate false positive links. The adjusted duplicate method curve and the absolute weight curve in Figure 8 agree closely, further corroborating the duplicate method as an accurate empirical method for estimating false positives.

Figure 8: Percentage of links likely to be false positives [1-PPV] by weight for an initial trial of pass 1, by Automatch® assigned weight (minimum weight 8), using both the duplicate method and the absolute weight method



3.3.4 Miscellaneous PPV estimates

3.3.4.1 Pass 2: Mortality records submitted to pass 1, versus those not submitted to pass 1

The high PPV for the Vitals-CAU pass (pass 2) was contrary to expectations, but seemed reasonable on further reflection. It was possible that the high PPV was mainly a function of mortality records being submitted to the pass that had *not* also been submitted to the meshblock pass (pass 1) - that is, those decedents with only a CAU on the SNZ Vitals file.

Of the 3552 mortality records involved in MP pairs for the Vitals-CAU pass, 2102 had not been submitted to the first pass. The PPV by the duplicate method for these 2102 MP pairs was 90.6%. Conversely, the PPV by the duplicate method for the remaining 1450 MP pairs was 90.9%. Therefore, there was no reason to separate the Vitals-CAU pass into those mortality records submitted to pass 1, and those not.

3.3.4.2 Pass 1: Tolerance of ±1 for day, month and year of birth

Observations of the NHI and NMDS day, month and birth of year variables prior to the record linkage suggested that a +/-1 disagreement on these variables for the same decedent was relatively common, and more likely than greater disagreements. Thus, a tolerance of +/-1 on these variables was allowed in the record linkage assuming that the 'amount' of disagreement with the census records might be similar.

Table 15: PPV estimates by the duplicate method for the +/-1 tolerances for day, month, and year of birth, for the first pass of the final match-run.

Tolerance	MP pairs	DA pairs	E[FP]	PPV
+/- 1 dd	172	43	23	87%
+/- 1 mm	139	27	14	90%
+/- 1 yyyy	121	3	2	99%
+/- 1 total	432	73	38	91%

Results in Table 15 suggests that the PPV for those records disagreeing by +/- 1 (NMDS and NHI agree but census disagrees with both NHI and NMDS by +/-1) varied by whether the partial agreement was for day, month, or year of birth. A partial agreement of +/-1 for year of birth resulted in a high PPV estimate (99%), but partial agreement on day (87%) and month of birth (90%) resulted in a lower PPV estimate, comparable with the PPV estimates for the CAU passes.

3.3.4.3 Pass 1: Disagreement on country of birth

It was uncertain a priori how accurate the country of birth matching variable would be. For pass 1 of the final match-run, there were 199 MP pairs with a disagreement on country of birth. Of these, 15 were estimated by the duplicate method to be false positives, or a PPV of 93%. Whilst numbers were small, there was some indication that disagreements between New Zealand, UK and Australia as the country of birth were less likely to result in the link being a false positive than disagreements between other countries. This supported the clerical review rules that allowed an 'acceptable' disagreement on country of birth.

3.3.5 Concluding comments on the PPV estimates in this research

Both the exact and duplicate method probably mildly overestimated the PPV in this research. The chance method probably overestimated the PPV as MPROB on the final match-run strategy moved the links to slightly higher weights, perhaps making the cut-off of 30.0 on the meshblock pass and 23.0 on the CAU passes fall a little below the ideal 'exact' match cut-offs. The duplicate method probably slightly overestimated estimated the PPV as it disregards false positives occurring from the incorrect retention of the highest weight scoring link when one mortality record links with more than one census record (MP pair with one or more DA pairs).

Taking all the above into account, the best PPV estimates by pass might reasonably be:

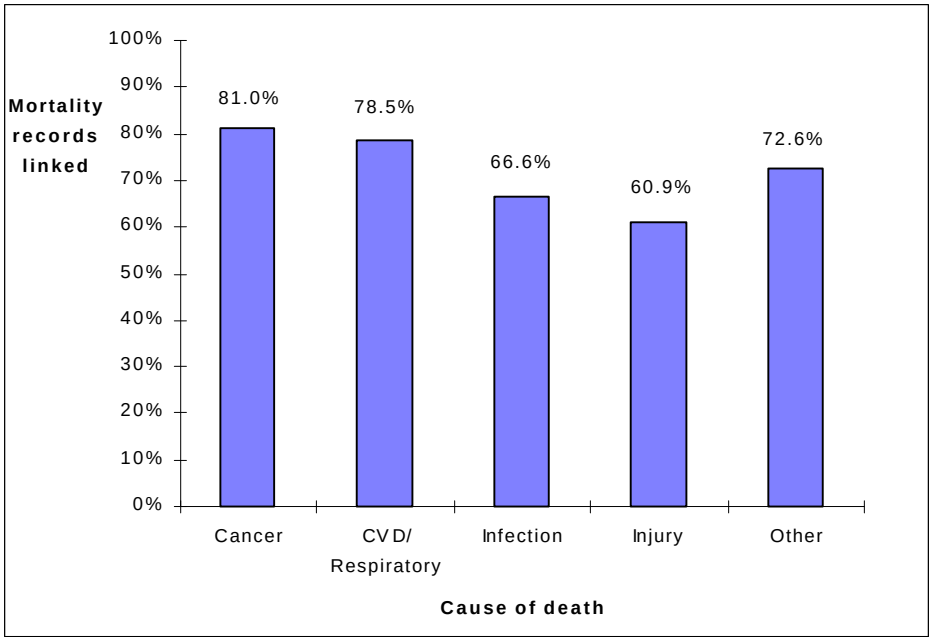
- 99.5% for the meshblock pass
- 90% for the Vitals-CAU pass
- 88% for the post-CAU pass
- 86% for the pre-CAU pass
- 83% for the NHI-CAU pass
- 80 to 90% for the clerical review of the meshblock and Vitals-CAU pass

(passes 6, 7, and 8).

Chapter 4: Results – analysis of bias

76.6% of the eligible mortality records were linked to a census record. This percentage varied by broad grouping of death as shown in Figure 9. The objective of this section on the analysis of bias in the record linkage is to determine the difference in linkage success by demographic and socio-economic variables, particularly the latter to allow quantification of what becomes a follow-up bias by socio-economic factors in the cohort analysis. Results are presented under two broad headings: stratified analyses and regression analyses. The methods used are described in the Methods Section (pages 63 to 63). Most analyses are for all deaths combined, although bias by specific cause of death are also investigated.

Figure 9: Percentage of mortality records linked by cause of death



4.1 Stratified analyses

Univariate and stratified results are presented by heading of demographic and socio-economic variable of interest.

4.1.1 Month following the census

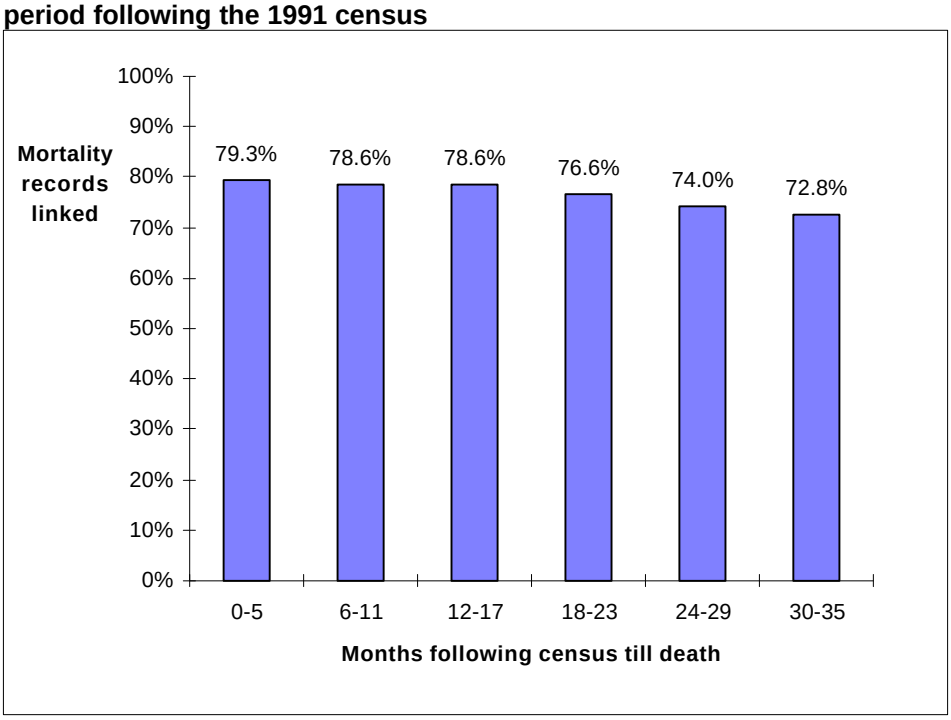
The mortality records were aggregated by date of death in six month periods following the 1991 census: 0-5, 6-11, …., and 30-35 months. Results are shown in Table 16, and Figure 10. The percentage of mortality records linked declined from 79.3% for the first six month period to 72.8% in the last six month period, or a relative risk of 0.92 for the last six months compared with the first six months. The percentage linked in the first six months was less than expected, but the percentage linked in the last six months was greater than expected. The reason for this may be that miscoding or incorrect information on either the census or mortality file caused much of the inability to link records (hence the approximately 20% of mortality records that could not be linked in the first six months), and that mobility was a lesser problem than expected for all deaths as reflected by the gradual fall-off in linkage over the three years.

Table 16: Mortality records (all deaths) linked by length of time (in six month periods) between the 1991 census and death

		Six month periods following the 1991 census						
Measure	Strata	0-5	6-11	12-17	18-23	24-29	30-35	Total
Number linked †		5271	5040	5628	5157	5418	5121	31635
Total decedents		6647	6413	7161	6735	7320	7034	41310
Percentage linked		**79.3%**	**78.6%**	**78.6%**	**76.6%**	**74.0%**	**72.8%**	**76.6%**
Relative risk		**-**	**0.99**	**0.99**	**0.97**	**0.93**	**0.92**	
				Age				
Percentage linked	0-14	73.5%	73.3%	70.6%	67.6%	65.8%	60.0%	69.0%
	15-24	63.0%	56.9%	54.4%	48.5%	50.5%	49.0%	53.7%
	25-44	72.6%	67.4%	67.3%	65.1%	59.0%	59.6%	65.1%
	45-64	80.0%	81.4%	79.2%	77.9%	74.6%	73.6%	77.6%
	65-74	82.1%	81.4%	82.9%	81.1%	78.7%	77.9%	80.6%
Relative risk	0-14	-	1.00	0.96	0.92	0.89	0.82	
	15-24	-	0.90	0.86	0.77	0.80	0.78	
	25-44	-	0.93	0.93	0.90	0.81	0.82	
	45-64	-	1.02	0.99	0.97	0.93	0.92	
	65-74	-	0.99	1.01	0.99	0.96	0.95	
				Ethnic group				
Percentage linked	Maori	62.4%	63.0%	64.7%	63.1%	63.3%	63.8%	63.4%
	Pacific	69.6%	64.6%	53.9%	56.4%	52.3%	53.5%	57.7%
	Rest	81.5%	80.4%	80.9%	78.5%	75.8%	74.5%	78.5%
	Not Spec	81.7%	86.5%	82.1%	83.3%	82.1%	70.6%	81.9%
Relative risk	Maori	-	1.01	1.04	1.01	1.01	1.02	
	Pacific	-	0.93	0.77	0.81	0.75	0.77	
	Rest	-	0.99	0.99	0.96	0.93	0.91	
	Not Spec	-	1.06	1.00	1.02	1.00	0.86	
				Rural - urban				
Percentage linked	Urban	80.9%	80.0%	80.0%	77.4%	75.0%	73.8%	77.8%
	Rural	69.2%	70.0%	69.0%	70.8%	67.5%	65.5%	68.6%
Relative risk	Urban	-	0.99	0.99	0.96	0.93	0.91	
	Rural	-	1.01	1.00	1.02	0.98	0.95	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.

Figure 10: Percentage of all mortality records linked to a census record by six month period following the 1991 census

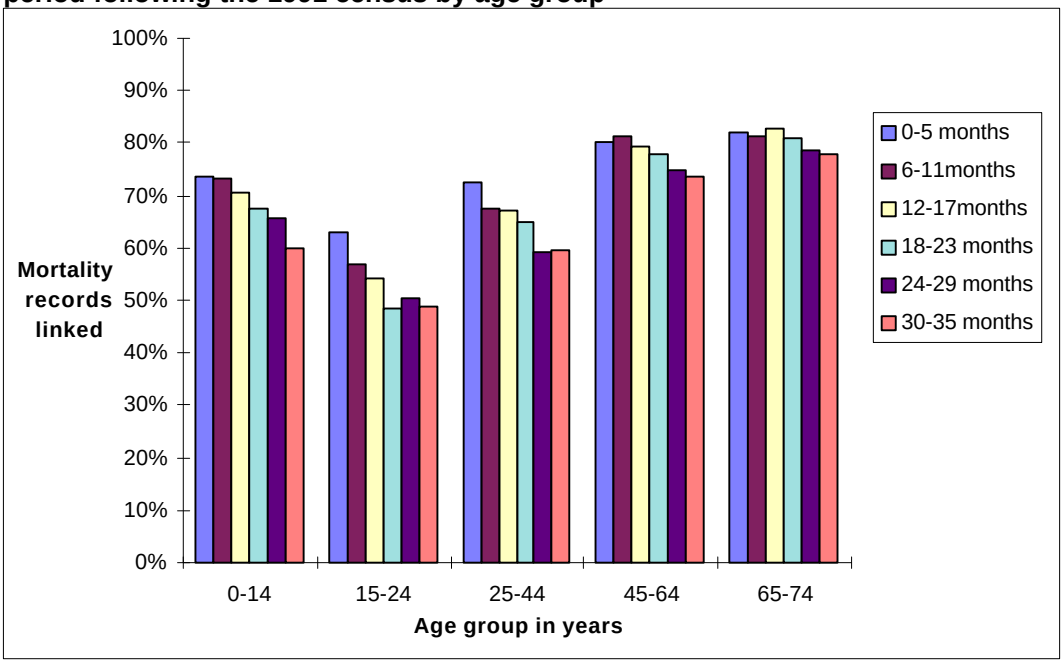


The linkage by six month period following the census varied by age group (Table 16 and Figure 11). The percentage linked to a census record declined more over time for decedents aged up to 44 years than for 45 to 74 year olds. This variation is presumably a function of increased residential mobility for younger people. But note that only two thirds of 15 to 24 year old decedents could be linked in the first six months, suggesting again that mobility was perhaps not the major reason for failure to link records.

The linkage by six month period following the census also varied by ethnic group and rurality as shown in Table 16. There was a tendency for the percentage of Maori and rural decedents linked to a census record to be comparatively stable over time following the census, compared to a decline for other strata. The lack of decline over time for rural decedents may be a function of improved geocoding (meshblock assignment) for rural addresses from 1991 to 1994.

There was no apparent variation by sex and socio-economic status (NZDep91 and NZSEI occupational class) for linkage by six month period following the census.

Figure 11: Percentage of all mortality records linked to a census record by six month period following the 1991 census by age group



4.1.2 Sex

Results of the linkage by sex are shown in Table 17. The small decreased percentage of male decedents linked with a census record compared to female decedents was consistent across strata of all other variables, except one age group - 61% of 25-44 year old male decedents were linked with a census record, whereas 72% of 25-44 year old females were linked. The small male to female discrepancy in linkage success was similar across broad groupings of cause of death, except injury with 59.0% of males being linked to a census record compared to 67.2% of females - this was probably a function of varying age by sex for injury deaths.

Table 17: Mortality records linked by sex

	Sex		
Measure	Male	Female	Total
Number linked †	19098	12534	31635
Total mortality records	25222	16088	41310
Percentage linked	75.7%	77.9%	76.6%
Relative risk	0.97	-	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.

4.1.3 Age

The mortality records were aggregated into both five year age groups by age at the census (0-4, 5-9, …., 70-74), and by five larger age groups (0-14, 15-24, 25-44, 45-64, 65-74). Figure 12 shows the percentage linked by five year age group.

Figure 12: Percentage of mortality records linked by five year age group

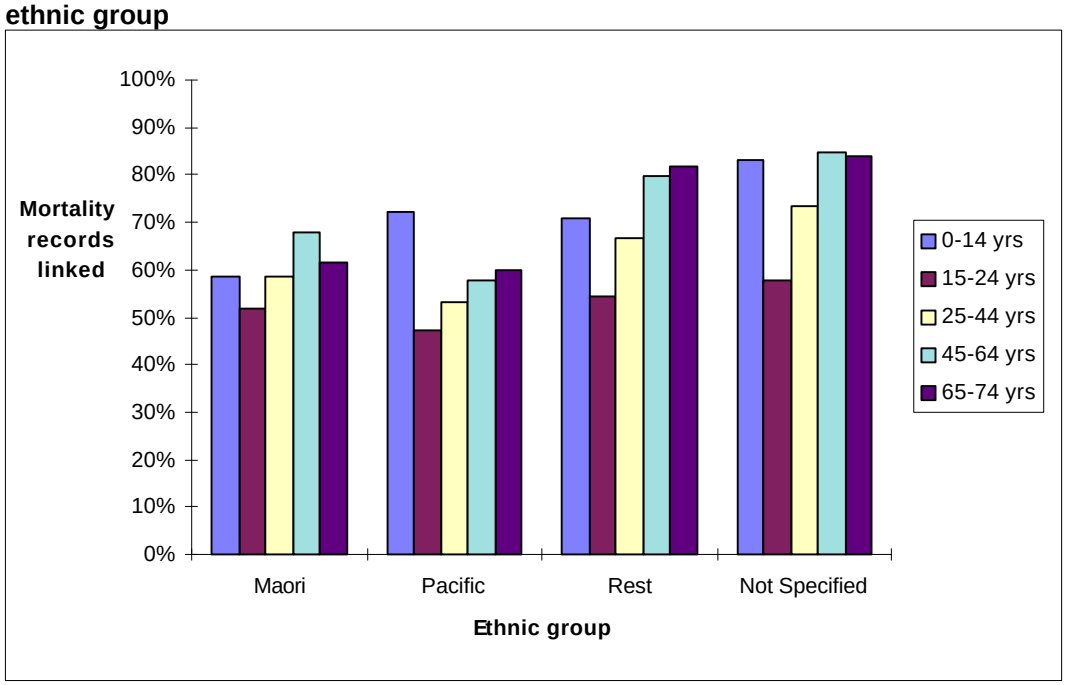
Results by larger age groups are shown in Table 18. The percentage of mortality records linked was greatest for 65-74 year olds (80.6%), and least for 15-24 year olds (53.7%). The linkage by age group arguably varied by ethnic group (Figure 13), with a shallower 'U' shaped distribution for Maori (and possibly Pacific people, although strata numbers are small). This interaction of age and ethnic group is more easily seen in the next section when ethnic group is presented stratified by age.

Table 18: Mortality records linked by age group on census night

		Age groups (years)					
Measure	Strata	0-14	15-24	25-44	45-64	65-74	Total
Number linked †		573	948	2739	11583	15789	31632
Total mortality records		831	1765	4207	14921	19586	41310
Percentage linked		69.0%	53.7%	65.1%	77.6%	80.6%	76.6%
Relative risk		0.86	0.67	0.81	0.97	-	
		Ethnic group					
Percentage linked	Maori	58.5%	51.7%	58.5%	68.0%	61.7%	63.4%
	Pacific	72.2%	47.4%	53.2%	57.8%	60.0%	57.7%
	Rest	70.8%	54.5%	66.8%	79.9%	82.0%	78.5%
	Not Spec	83.3%	57.9%	73.5%	84.7%	84.1%	81.9%
Relative risk ‡	Maori	0.95	0.84	0.95	1.10	-	
	Pacific	1.20	0.79	0.89	0.96	-	
	Rest	0.86	0.66	0.81	0.98	-	
	Not Spec	0.99	0.69	0.87	1.01	-	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.
‡ Reference category is 65-74 year olds.

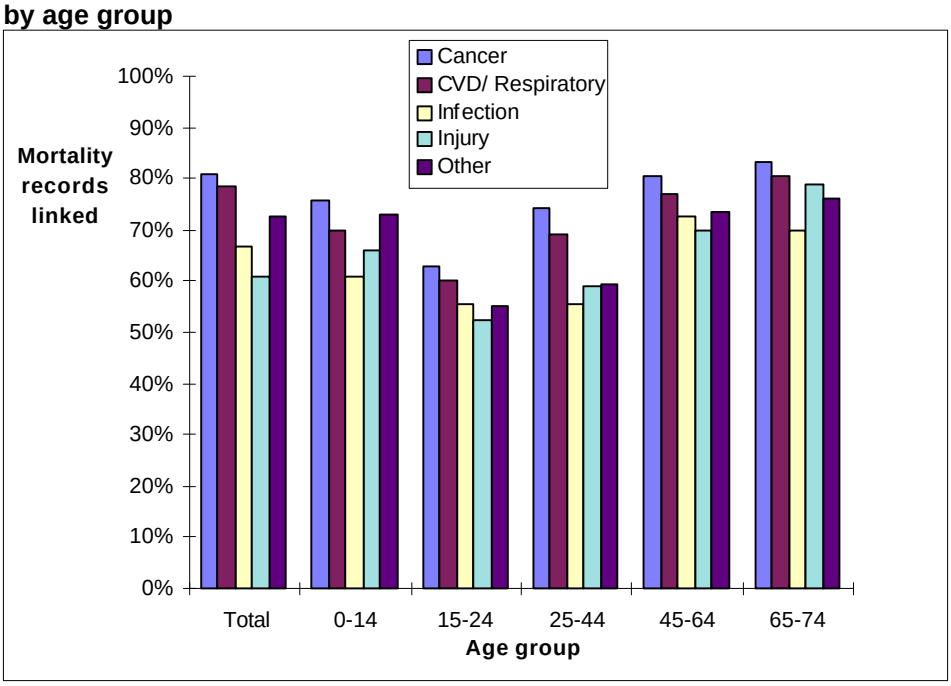
Figure 13: Percentage of mortality records linked to a census record by age group by ethnic group



Stratification of age by other demographic and socioeconomic variables demonstrated no obvious confounding or effect modification of the underlying association of age with linkage to a census record for all deaths combined.

Figure 13 shows that variation in linkage rates by cause of death was partly due to age. For example, the discrepancy between the percentage of injury deaths linked is less within age groups than for the total population. Perhaps a third to a half of the discrepancy in linkage rates by cause of death is explained by age group.

Figure 14: Percentage of mortality records linked to a census record by cause of death by age group



4.1.4 Ethnic group

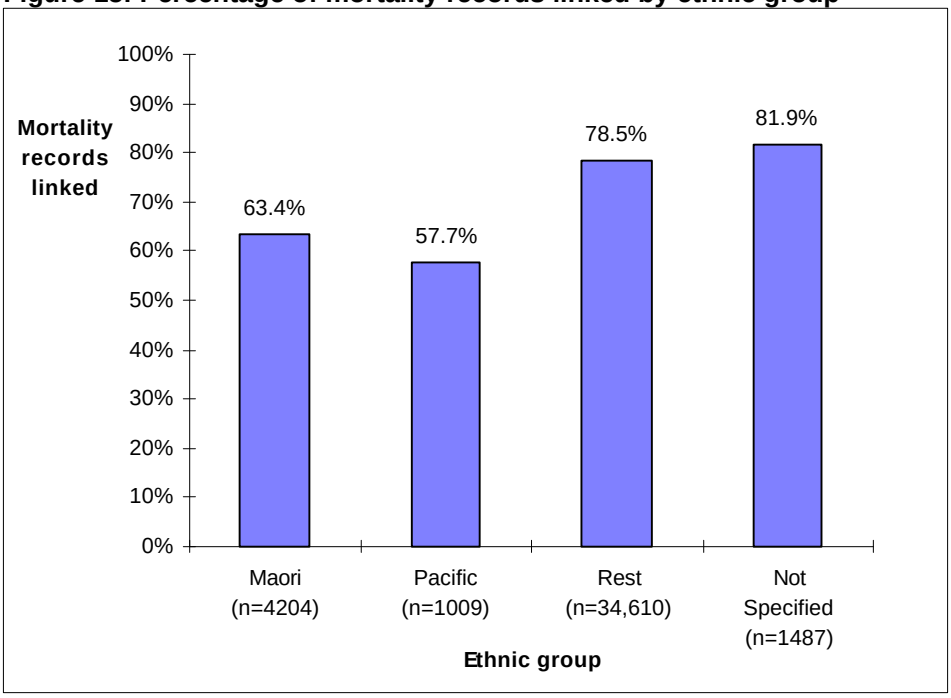
The results presented here are for aggregated ethnic groups based on that recorded on the NHI File (Table 19). Excluding the not specified, the percentage of mortality records linked was greatest for the 'Rest' (81.9%; those people with a specified ethnic group that is neither Maori nor Pacific, and will mainly be New Zealand Europeans), and lower for Maori (63.4%) and Pacific people (57.7%; Figure 15).

Table 19: Mortality records linked by ethnic group

		Ethnic group				
Measure	Strata	Maori	Pacific	Rest	Not Spec	Total
Number linked †		2667	582	27168	1218	31635
Total mortality records		4204	1009	34610	1487	41310
Percentage linked		63.4%	57.7%	78.5%	81.9%	76.6%
Relative risk ‡		0.81	0.73	-	1.04	
		Age				
Percentage linked	0-14	58.5%	72.2%	70.8%	83.3%	69.0%
	15-24	51.7%	47.4%	54.5%	57.9%	53.7%
	25-44	58.5%	53.2%	66.8%	73.5%	65.1%
	45-64	68.0%	57.8%	79.9%	84.7%	77.6%
	65-74	61.7%	60.0%	82.0%	84.1%	80.6%
Relative risk ‡	0-14	0.83	1.02	-	1.18	
	15-24	0.95	0.87	-	1.06	
	25-44	0.88	0.80	-	1.10	
	45-64	0.85	0.72	-	1.06	
	65-74	0.75	0.73	-	1.03	
		NZDep91 quintile				
Percentage linked	1	63.3%	44.4%	82.6%	86.2%	82.4%
	2	67.1%	64.3%	81.8%	85.1%	81.3%
	3	64.4%	64.5%	81.2%	81.9%	80.1%
	4	65.3%	60.0%	79.7%	82.8%	77.9%
	5	67.7%	56.0%	77.3%	80.2%	74.2%
Relative risk ‡	1	0.77	0.54	-	1.04	
	2	0.82	0.79	-	1.04	
	3	0.79	0.79	-	1.01	
	4	0.82	0.75	-	1.04	
	5	0.88	0.73	-	1.04	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.
‡ Reference category is the 'Rest'.

Figure 15: Percentage of mortality records linked by ethnic group

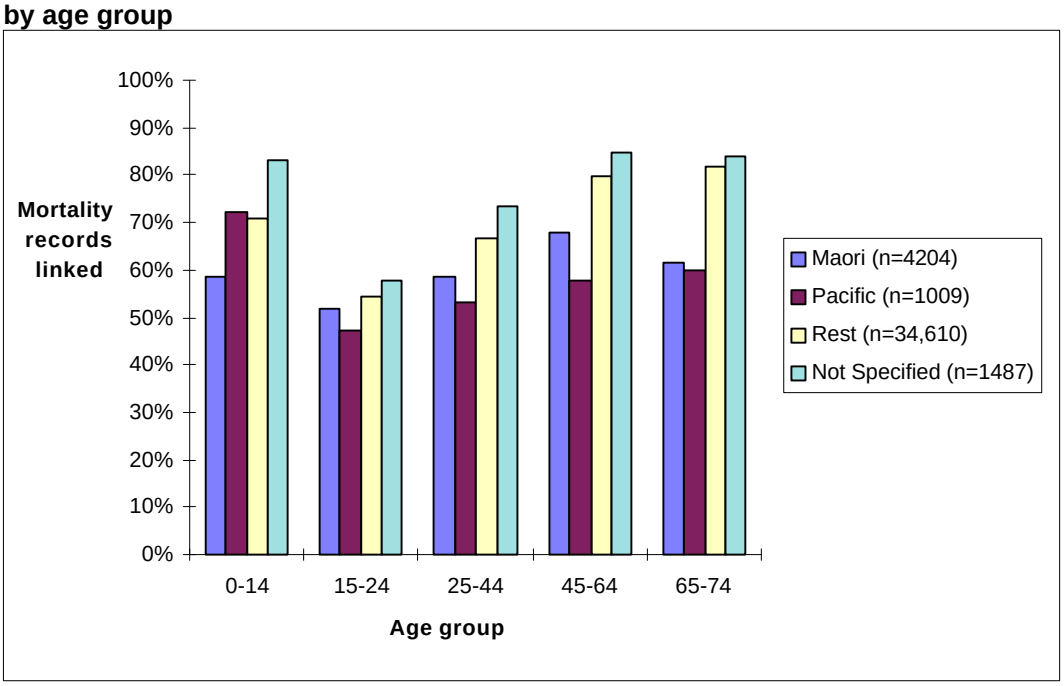


The linkage by ethnic group varied across age groups as shown in Table 19 and Figure 16. Within young adult age groups (15 to 44 years) there is less relative difference in the percentage of mortality records linked between Maori and the Rest. (Numbers for Pacific people and not specified are small for 15 to 44 year olds, limiting inference). That is, age group and ethnic group interact in their association with linkage to a census record.

Stratification by sex found no substantial variation in the underlying association of ethnic group with linkage.

There was no marked control of the association of ethnic group with linkage by NZDep91 quintile (Table 19), but there was some possible interaction of ethnic group and NZDep91 quintile that is more clearly presented in the subsequent section on NZDep91.There was no apparent variation across NZSEI categories.

Figure 16: Percentage of mortality records linked to a census record by ethnic group by age group



The pattern of variation in linkage rates by cause of death (Figure 9, page 97) was similar for each ethnic group.

4.1.5 Urban or rural usual residence

There was no apparent difference in the percentage of mortality records linked between main, secondary and minor urban areas, therefore they were combined into one urban group. Results for record linkage for urban versus rural are shown in Table 20. Rural mortality records were less likely to be linked with a census record than urban mortality records. Much of the reason for this difference is that rural mortality records are less likely to be assigned a meshblock than urban mortality records, due to difficulties geocoding rural addresses. Thus, when only mortality records that had a meshblock assigned are considered, the relative risk of linkage to a census record for rural compared to urban mortality records was 0.96, a two thirds reduction to the null from the non-stratified relative risk of 0.88 (Table 20). (Only 43% of the rural mortality records had a meshblock assigned compared to 96% of urban mortality records).

Table 20: Mortality records linked by urban or rural residence, including stratification by whether a meshblock was assigned

Measure	Meshblock assigned?	Urban - rural Urban	Rural	Total
Number linked †		28080	3537	31617
Total mortality records		36101	5154	41255
Percentage linked		77.8%	68.6%	76.6%
Relative risk		-	0.88	
Percentage linked	yes	78.9%	75.4%	
	no	51.0%	63.5%	
Relative risk	yes	-	0.96	
	no	-	1.25	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.

Beyond meshblock assignment, there was no obvious variation in the association of rurality with linkage when stratified by demographic and socio-economic variables.

4.1.6 Bias by RHA

Record linkage varied by regional health authority (RHA), with lower linkage rates in the Northern and Midland RHAs (73.3% and 74.8% respectively) than the Central and Southern RHAs (79.0% and 80.0%). Some of this variation was due to confounding by ethnic group (Table 21 and Figure 17). There was also apparent effect modification of the association of RHA with linkage by age (Table 21 and Figure 18), with a steeper gradient across RHAs for younger people - although this may in turn be due to confounding and/or effect modification by ethnic group.

The goal of the NZCMS is to determine socio-economic mortality gradients by sex, age and ethnic groups at a national level. Therefore, further investigation of regional effects is not pursued in this technical report. When the NZCMS moves onto a consideration of socio-economic mortality gradients *by region*, further analysis of bias by region will be required.

Table 21: Mortality records linked by RHA

		RHA				
Measure	Strata	Northern	Midland	Central	Southern	Total
Number linked †		9078	6561	8286	7695	31620
Total mortality records		12377	8767	10494	9617	41255
Percentage linked		73.3%	74.8%	79.0%	80.0%	76.6%
Relative risk ‡		0.96	0.98	1.03	1.04	
		Ethnic group				
Percentage linked	Maori	60.9%	63.8%	66.6%	66.2%	63.4%
	PI	57.2%	64.7%	59.5%	53.9%	57.7%
	Rest	76.5%	77.6%	79.8%	80.4%	78.5%
	Not Spec	66.7%	73.3%	85.0%	81.1%	81.9%
Relative risk ‡	Maori	0.96	1.01	1.05	1.04	
	PI	0.99	1.12	1.03	0.93	
	Rest	0.97	0.99	1.02	1.02	
	Not Spec	0.81	0.90	1.04	0.99	
		Age				
Percentage linked	0-14	62.5%	63.2%	72.5%	82.4%	69.0%
	15-24	46.5%	56.0%	53.6%	61.4%	53.7%
	25-44	62.5%	63.2%	69.6%	67.0%	65.1%
	45-64	74.3%	76.7%	79.9%	81.1%	77.6%
	65-74	78.3%	78.5%	82.7%	82.9%	80.6%
Relative risk ‡	0-14	0.91	0.92	1.05	1.19	
	15-24	0.87	1.04	1.00	1.14	
	25-44	0.96	0.97	1.07	1.03	
	45-64	0.96	0.99	1.03	1.04	
	65-74	0.97	0.97	1.03	1.03	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.
‡ The reference for the relative risk is the total (by strata).

Figure 17: Percentage of mortality records linked to a census record by RHA by ethnic group

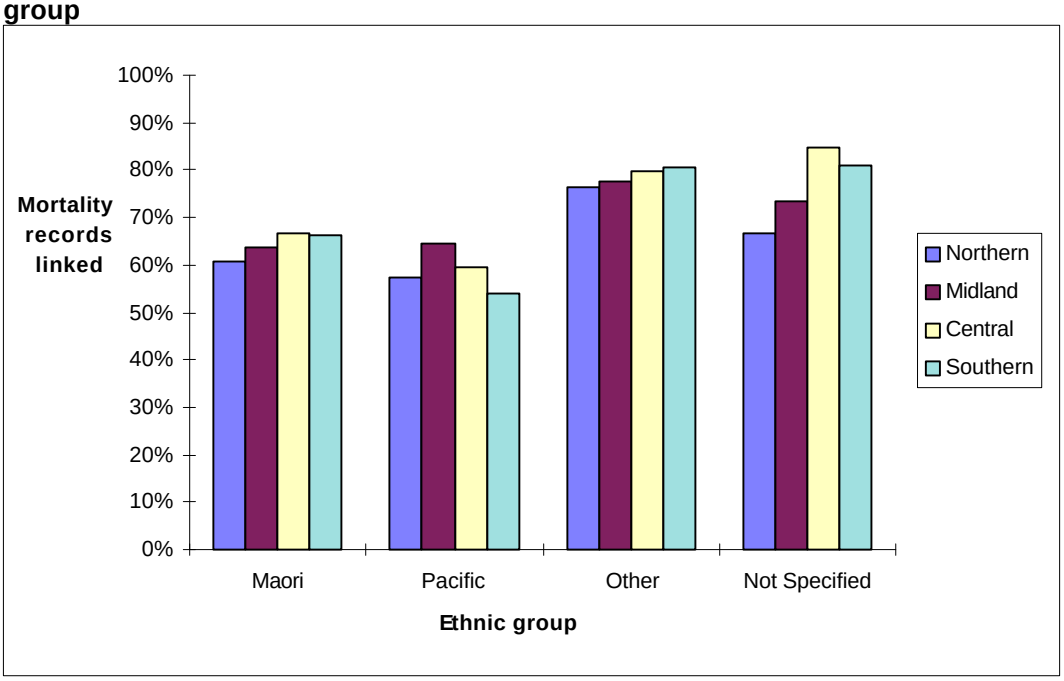
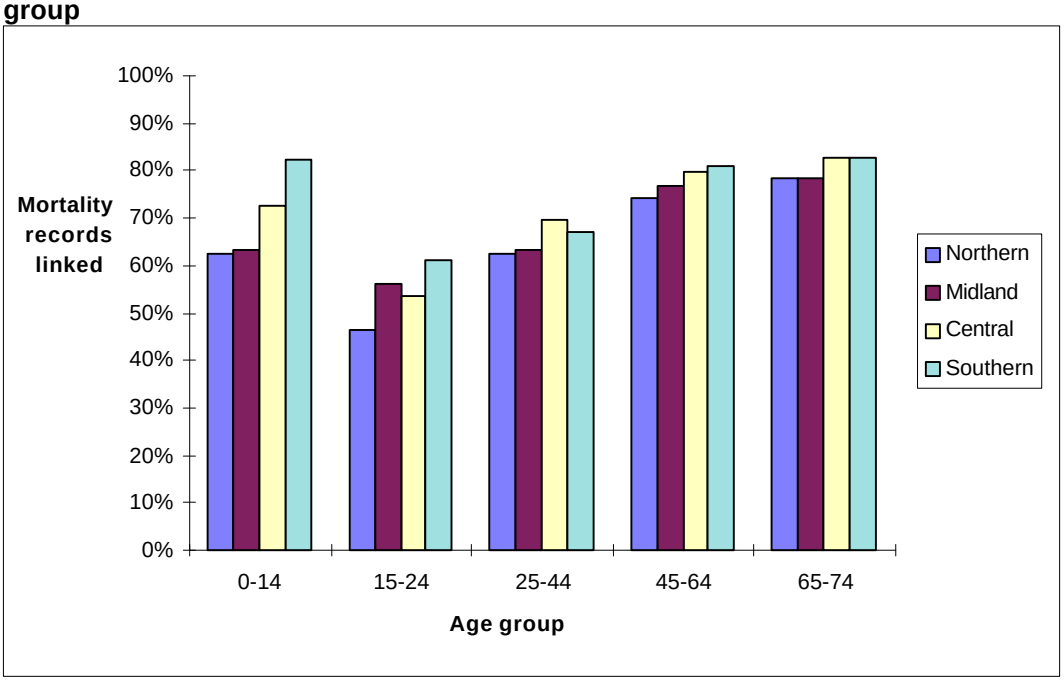


Figure 18: Percentage of mortality records linked to a census record by RHA by age group



4.1.7 NZDep91 small area deprivation

The percentage of mortality records linked to a census record was highest for the least deprived decile (82.8%), and lowest for the most deprived decile (71.6%; Table 22). Note that there is a gradual linear decline from decile 1 to decile 9, but then a larger drop from decile 9 (76.8%) to decile 10 (71.6%).

Table 22: Mortality records linked by NZDep91 decile

	NZDep decile										
Measure	1	2	3	4	5	6	7	8	9	10	Total
Linked	2427	2409	2475	2562	2742	2892	3135	3336	3567	3492	29037
Total	2932	2940	3023	3174	3418	3617	4024	4280	4644	4875	36927
% linked	82.8%	81.9%	81.9%	80.7%	80.2%	80.0%	77.9%	77.9%	76.8%	71.6%	78.6%
Relative risk	-	0.99	0.99	0.98	0.97	0.97	0.94	0.94	0.93	0.87	

Results of NZDep91 quintile stratified by ethnic group and NZSEI occupational class are shown in Table 23. Perhaps a third of the association of NZDep91 with linkage was explained by ethnic group.

The percentage of Maori mortality records linked to a census record was relatively stable across NZDep91 quintiles compared to the decline with increasing deprivation for other ethnic groups. This is likely to be a function of misclassification bias due to the distribution of Maori decedents' usual residence (and Pacific decedents') being strongly skewed to the more deprived small areas, compared to non-Maori, non-Pacific decedents. Consider the following simplified example. Assume that small area deprivation is a binary variable, that 80% of Maori decedents resided in deprived small areas (and therefore 20% of Maori decedents resided in non-deprived small areas), and that 50% of non-Maori, non-Pacific decedents lived in deprived small areas. Assume also that there was random misclassification of small area deprivation such that 20% of decedents truly residing in either deprived or non-deprived small areas were misclassified to the other. Then we would observe 68% of Maori decedents assigned to deprived small areas (94% correctly assigned, the other 6% actually misclassified Maori decedents living in non-deprived small areas), 32% of Maori decedents assigned

to non-deprived small areas (50% correctly assigned, 50% incorrectly assigned), and 50% of non-Maori, non-Pacific assigned to both deprived and non-deprived small areas (80% correctly assigned, and 20% incorrectly assigned in both instances). Note that this assignation of deprivation is highly correlated with the chance of being linked to a census record - small area deprivation is assigned on the basis of meshblocks, and having a meshblock assigned greatly increases the probability of being linked to a census record. If 90% of decedents correctly assigned to either deprived or non-deprived small areas were linked to a census record (by virtue of having a correct meshblock also assigned), but none of the decedents incorrectly assigned were linked to a census record (as their assigned meshblock was also incorrect), then we would observe:

- $0.94 \times 0.90 = 85\%$ of Maori decedents assigned as deprived being linked to a census record
- $0.50 \times 0.90 = 45\%$ of Maori decedents assigned as non-deprived being linked
- $0.80 \times 0.90 = 72\%$ of non-Maori, non-Pacific decedents linked for both those assigned as deprived and those assigned as non-deprived.

That is, because the NZDep91 score assignation and the probability of being linked to a census record are both dependent on the meshblock being assigned correctly, **and** the Maori decedents are skewed towards deprived small area residence, random misclassification of meshblock may mean that gradients in linkage across NZDep91 scores for Maori are more biased than that for non-Maori, non-Pacific decedents. Returning to the data presented in Table 23, it is probable that some random misclassification of meshblock for Maori will underestimate the percentage of Maori decedents linked who truly reside in a non-deprived small area relative to those who truly live in a deprived small area. Thus, it is likely that the lack of gradient in record linkage by NZDep91 quintile for Maori is partly a function of misclassification bias, not solely a function of a lack of an underlying gradient in linkage by socio-economic factors.

Not shown in Table 23 is the percentage linked by NZDep91 quintile by age group - there was little variation in the slope by age group, except for a possibly flatter slope for 65-74 year olds.

Table 23: Mortality records linked by NZDep91 quintile

		NZDep91 quintile					
Measure	Strata	1	2	3	4	5	Total
Number linked †		4836	5040	5634	6471	7059	29040
Total mortality records		5872	6197	7035	8304	9519	36927
Percentage linked		82.4%	81.3%	80.1%	77.9%	74.2%	78.6%
Relative risk ‡		-	0.99	0.97	0.95	0.90	
Relative risk #		1.11	1.10	1.08	1.05	-	
		Ethnic group					
Percentage linked	Maori	63.3%	67.1%	64.4%	65.3%	67.7%	63.4%
	Pacific	*	*	64.5%	60.0%	56.0%	57.7%
	Rest	82.6%	81.8%	81.2%	79.7%	77.3%	78.5%
	Not Spec	86.2%	85.1%	81.9%	82.8%	80.2%	81.9%
Relative risk #	Maori	0.94	0.99	0.95	0.96	-	
	Pacific	*	*	1.15	1.07	-	
	Rest	1.07	1.06	1.05	1.03	-	
	Not Spec	1.08	1.06	1.02	1.03	-	
		NZSEI category%					
Percentage linked	1	82.7%	86.5%	85.9%	78.9%	73.4%	82.2%
	2	82.2%	84.3%	83.2%	82.8%	75.6%	81.9%
	3	83.6%	81.3%	80.7%	78.9%	79.7%	81.0%
	4	82.7%	82.5%	81.6%	83.4%	77.9%	81.5%
	5	83.5%	82.9%	80.7%	79.5%	76.0%	79.7%
	6	70.4%	76.5%	73.6%	75.7%	74.7%	74.6%
	Farmers	82.1%	76.6%	81.3%	74.4%	68.9%	76.8%
	No occup	80.1%	79.5%	77.7%	76.3%	73.1%	76.6%
Relative risk ‡	1	1.00	1.05	1.04	0.95	0.89	
	2	1.00	1.03	1.01	1.01	0.92	
	3	1.00	0.97	0.97	0.94	0.95	
	4	1.00	1.00	0.99	1.01	0.94	
	5	1.00	0.99	0.97	0.95	0.91	
	6	1.00	1.09	1.05	1.08	1.06	
	Farmers	1.00	0.93	0.99	0.91	0.84	
	No occup	1.00	0.99	0.97	0.95	0.91	

† Random rounded to three digit intervals as per standard SNZ protocol. Total not rounded.
‡ Reference category is quintile 1.
# Reference category is quintile 5 due to small numbers in quintile 1 for Maori and Pacific people.
* Suppressed data by either SNZ protocol (small cell sizes).
% For decedents age 25-74 on census night, and dying in the second and third year after the census. As described in Appendix 1, the NZSEI occupational class groupings used are slightly different than those proposed by Davis et al (1997).[30]

Table 23 also shows the association of NZDep91 with linkage within NZSEI occupational class categories, for the 24,009 decedents aged 25-74 on census night

and dying in the second and third years after census night. (Deaths within the first year of follow-up were discarded as NZSCO90 codes on 1991 mortality data were not correct (personal communication, Jim Fraser, Manager, NZHIS, June 1999).) The association of NZDep91 with linkage remains within NZSEI occupational class categories, albeit often a reduced association compared to that non-stratified for NZSEI occupational class. The reversed slope within occupational class 6 should be treated with caution, as there were only 69 decedents in the NZDep91 reference cell, the least deprived quintile of small areas. The exact inter-relationships of NZDep91 and NZSEI scores will be more complex than presented here due to the variation in both probability of assignment of a NZSEI score, and variation of the value of any assigned score, by strata of sex, age and ethnic group - this was explored more with regression models in the subsequent section. However, it was plausible that some of the effect of NZDep91 is 'explained' by NZSEI occupational class category.

The pattern of variation in linkage rates by cause of death (Figure 9, page 97) was similar for each NZDep91 quintile.

4.1.8 NZSEI occupational class

The classification of occupational class from NZSEI scores used in this project is slightly different to that proposed by Davis et al for reasons discussed in Appendix 1. Instead, the boundary between occupational class 1 and 2 has been shifted, and farmers are removed from occupational class 6 to make a distinct 'occupational class'. For occupational class analyses, all deaths in the first year of follow-up were discarded due to incorrect data. During 1991 the occupational codes used for mortality were coded initially to NZSCO68 codes (New Zealand Standard Classification of Occupations, 1968 version). Subsequently these NZSCO68 codes were recoded as NZSCO90 codes – thus the NZSCO90 codes were not derived directly from the occupation written on the death registration form (personal communication, Jim Fraser, Manager, NZHIS, 1999). The NZSEI index requires NZSCO90 codes. Initial examinations of results by NZSEI occupational class derived from mortality data in this project demonstrated implausible findings for the first year of deaths. It was concluded that the concordance of NZSCO68 and NZSCO90 was not good enough to use occupational codes from

1991 mortality data. Occupational codes for 1992-1994 inclusive were directly coded to NZSCO90 codes.

Results by NZSEI occupational class category are shown in Table 24, and show a similar pattern to that seen with NZDep91 quintiles – a moderate drop in linkage success from occupational class 1 (83.6%) to 5 (77.5%), then a notable further drop from occupational class 5 and 6 (71.7%). Stratification by ethnic group reduced the gradient between NZSEI occupational classes 1 to 5 for Maori and the 'Rest', but a notable drop was still apparent from occupational class 5 to 6. Caution is required interpreting trends for Pacific people due to small numbers. Likewise, cell sizes are relatively small for Maori in occupational classes 1, 2 and 3 (26, 72, and 84 respectively).

Stratification by NZDep91 resulted in no obvious change in the gradient of linkage by NZSEI occupational class – indeed the within NZDep91 quintile associations of occupational class with linkage are somewhat erratic (Figure 19). All cells, except one, formed by cross-classifying NZDep91 quintile with occupational classes 1 to 6 had greater than 100 decedents.

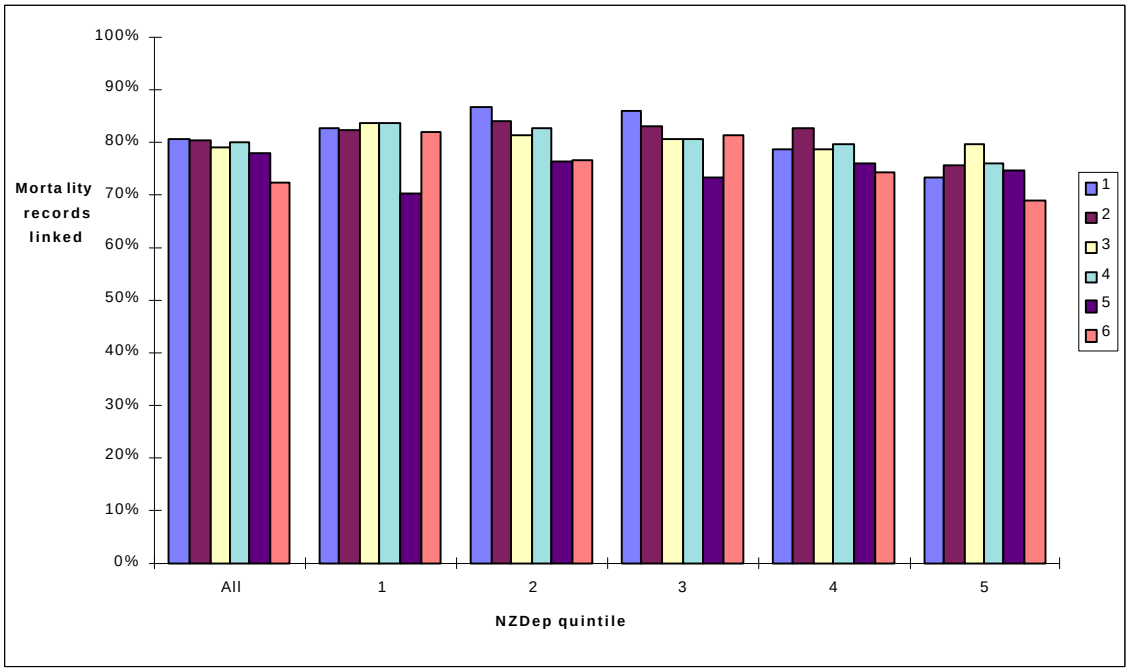
Table 24: Mortality records linked by NZSEI occupational class for 25-74 year olds †

		NZSEI occupational class							
Measure	Total	1	2	3	4	5	6	Farmers	No occ
Number linked	20394	2025	1932	2874	4413	3453	1482	2196	12684
Total records	26573	2421	2379	3633	5541	4458	2067	3036	16941
Percentage linked	78.9%	83.6%	81.2%	79.1%	79.6%	77.5%	71.7%	72.3%	74.9%
Relative risk		1.05	1.02	0.99	1.00	0.97	0.90	0.91	0.94
				Ethnic group					
Percentage linked	Maori	69.2%	75.0%	75.0%	69.4%	69.1%	60.2%	59.5%	63.2%
	Pacific	--------	52.2%	--------	62.1%	53.8%	68.4%	52.2%	51.9%
	Rest	80.5%	81.3%	79.2%	80.7%	80.1%	75.7%	73.1%	77.8%
	Not Sp	92.0%	77.8%	83.2%	83.3%	83.7%	65.9%	85.7%	81.6%
Relative risk ‡	Maori	1.00	1.08	1.08	1.00	1.00	0.87	0.86	0.91
	Pacific	---------	0.84	---------	1.00	0.87	1.10	0.84	0.84
	Rest	1.00	1.01	0.98	1.00	0.99	0.94	0.91	0.96
	Not Sp	1.10	0.93	1.00	1.00	1.00	0.79	1.03	0.98

† For decedents age 25-74 on census night, and dying in the second and third year after the census. As described in Appendix 1, the NZSEI occupational class groupings used are slightly different than those proposed by Davis et al (1997).[30] Due to small numbers, occupational classes 1-3 are aggregated for Pacific people.

‡ Occupational class 4 is reference category

Figure 19: Percentage of mortality records linked to a census record by NZSEI occupational class (excluding farmers) by NZDep91 quintile



4.1.9 Pass of the record linkage stratified by demographic and socio-economic variables

The analysis of bias by separate pass (of the eight passes of the final match-run) was conducted by SNZ as it was considered too great a privacy risk to be done by non-SNZ staff in the Data Laboratory. What follows is therefore a qualitative summary only.

4.1.9.1 Time following census

With each successive six month period between the census and death, pass 3 (post-CAU) was responsible for a steadily increasing percentage of the total number of mortality records linked. This was to be expected given that the reason for including the post-CAU pass was to 'find' people who had moved between census and death, and the chance of moving would increase with increasing time between the census and death. Likewise, but to a lesser extent, pass 4 (pre-CAU) increased its relative yield with increasing time following the census.

4.1.9.2 Age

Pass 3 (post-CAU), pass 4 (pre-CAU), and pass 5 (NHI-CAU) were relatively more important for young adults. This was to be expected given that young adults are more likely to move residence than other age groups.

4.1.9.3 Ethnic group

Pass 2 (Vitals-CAU) was relatively important for Maori. Passes 2 to 8 all had a lower relative yield for Pacific people. Two possible reasons for this latter finding are: a) Pacific people may have been recent immigrants to New Zealand without previous hospitalisation events (passes 3 to 5); and b) there may be poorer linking of health events by the NHI number for Pacific people compared to other ethnic groups(passes 3 to 5 again; the distinction between first and surnames among Pacific people is not as clear cut as for European New Zealanders, often resulting in one Pacific person having two hospital numbers for the two different orderings of names).

4.1.9.4 Rurality

Pass 2 (Vitals-CAU) was relatively important for rural decedents. This was as expected given that only about half of rural decedents had a meshblock assigned, making pass 2 the first opportunity for linkage for many rural decedents.

4.1.9.5 NZDep91

Pass 2 (Vitals-CAU) was better at higher deprivation, pass 1 at lower deprivation.

4.1.9.6 NZSEI

As with NZDep91, pass 2 (Vitals-CAU) was better for lower occupational classes, pass 1 for higher occupational classes.

4.2 Regression analyses

Results presented in the previous stratified analyses suggested that cohort analyses at a national-level should be conducted and presented separately by strata of sex, age, and ethnic group. Thus the objective in this Section is to quantify the follow-up bias by socio-economic factors *within sex, age and ethnic group strata*. Accordingly these three variables are included as covariates in the regression models. It was not the objective at this stage to quantify the follow-up bias by sub-national regions (eg RHA, rural-urban).

The multiple regression analyses of follow-up bias are presented under subheadings 4.2.1 to 4.2.5. Subheading 4.2.1 is preliminary to the remaining four that consider socio-economic follow-up bias directly, being multiple regression analyses of the bias by: age, sex, and ethnic group; and time period between the census and death controlling for age, sex, and ethnic group. The justification for these preliminary analyses were:

- to determine the under-representation of deaths by demographic strata, permitting adjustment of future observed stratum specific mortality risks in the cohort analysis
- to provide reference models for the subsequent regression analyses investigating bias by socio-economic measures.

Subheadings 4.2.2 to 4.2.5 consider multiple regression analyses of follow-up bias by socio-economic measure:

- NZDep91 small area deprivation (n=36,927 mortality records with a NZDep91 score, 89.4% of all mortality records)
- whether occupation was recorded on the BDM28 (all n=40,479 mortality records aged 15 years and over)
- NZSEI occupational class (n=15,760 mortality records aged 25-74 years with an occupation recorded; n=13,701 for males, n=1,884 for females)
- NZDep91 score and NZSEI occupational class, considered together (n=14,133 mortality records aged 25-74 year olds with both a NZDep91 score and an occupation recorded; n=12,249 for males, n=1,884 for females).

Given the different interpretation of occupation between sexes, the last three analyses were conducted separately by sex.

Together, these four different investigations allow comment on follow-up bias by socio-economic status for cohort analyses of *all-cause mortality* for deaths linked in the *full three year period following the census*. But some analyses may be restricted to deaths occurring in a shorter period following the census, and there will also be separate analyses by cause of death. Therefore, subsidiary regression models were fitted: first a model including time period following the census as a covariate, and second models separately by major causes of death. Both cause of death and time period are significant predictors of linkage (see Section 4.1, and subheadings, on stratified analyses). The specific question here for the subsidiary models was "does follow-up bias by socio-economic factors vary by either: a) cause of death, or b) time period?"

4.2.1 Demographic factors and time period

4.2.1.1 All cause mortality: sex, age and ethnic group

Figure 20 presents the percentage of mortality records linked by sex, age, and ethnic group, i.e. completely stratified by demographic factors.

A complete log-linear risk model for linkage was specified for all 41,310 mortality records, including:

- the single second order interaction term, [sex]×[age group]×[ethnic group]
- the three first order product terms, [sex]×[age group], [sex]×[ethnic group], and [age group]×[ethnic group]
- and the three main effects.

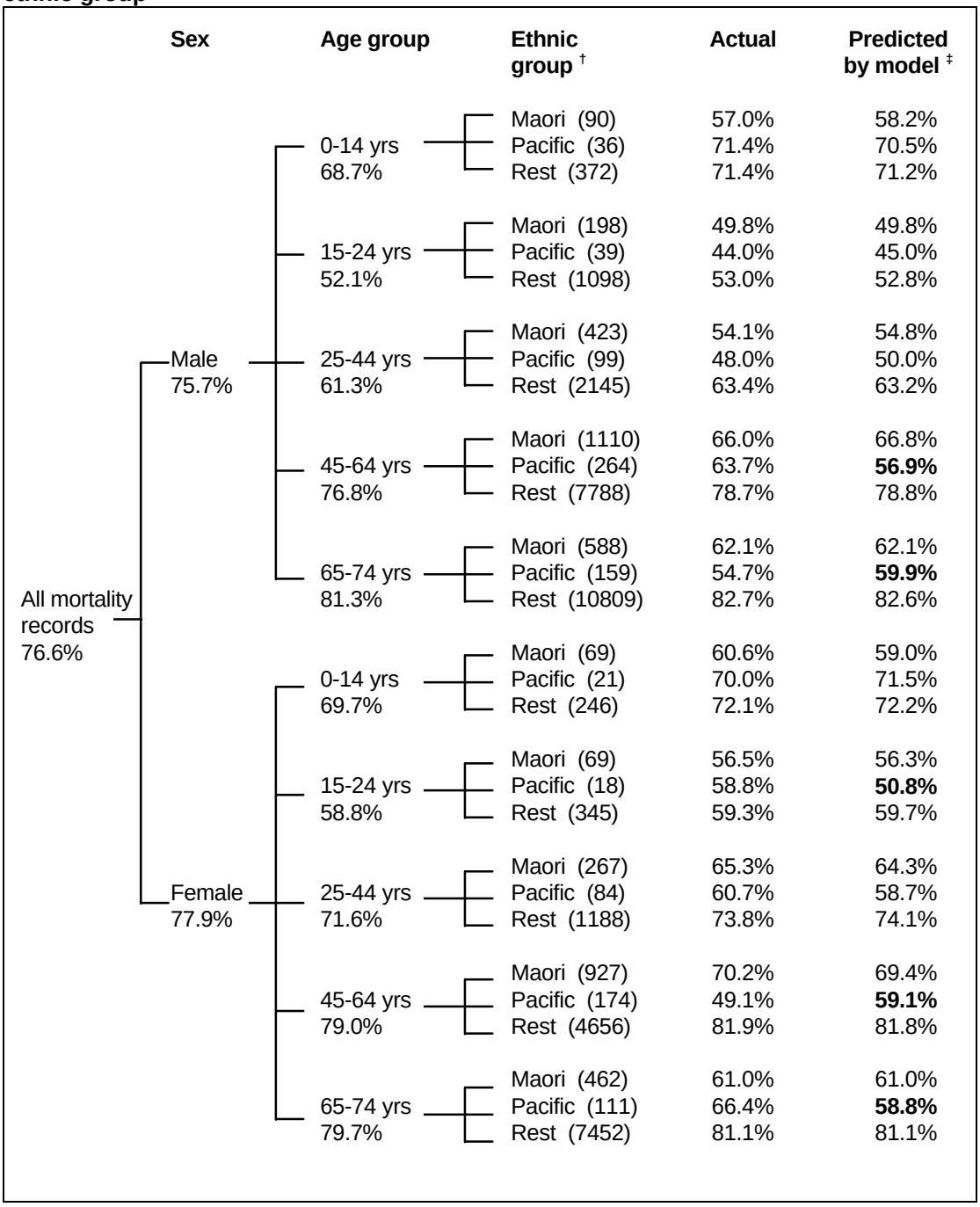
The second order interaction product was included in this model for completeness given that only three main effects were being considered, and because results in Section 4.1 indicated two first order interactions of [sex] × [age], and [age] × [ethnic

group], but a possible second order interaction involving all three had not been investigated.

Using the criteria for backwards elimination strategy on page 67 for model selection alone, the suggested final model was simply one with just the second order interaction term. The p value for this interaction product was 0.033 on Wald's Type III test for the complete model specified above. That is, the suggested model was equivalent to considering all 2×5×3 = 30 sex, age, and ethnic group strata individually - nothing was gained from modeling over a simple stratification by all three variables, as shown in Figure 20.

If a backwards elimination strategy started with just the three first order interaction products and the main effects, [sex] × [age] and [age] × [ethnic group] were statistically significant on the Wald's Type III test (p=0.0001 for both) but not [sex]×[ethnic group] (p=0.4823). This confirmed the results in Section 4.1 indicating that the former two first order interactions were important but not the latter. The estimated percentage of mortality records linked by strata using a log linear risk model with just the two interaction products ([sex]×[age group], and [age group]×[ethnic group]) are shown in the far right column of Figure 20. These estimated percentages differed from the actual stratum percentage for Pacific people only, most notably for older Pacific people (percentages in bold in Figure 20). The number of Pacific males aged 45-64 estimated to be linked to a census record by the model was 18 less than the actual number, but it was 17 more for Pacific females aged 45-64. Conversely, for Pacific males aged 65-74 the model estimate was 8 greater, and that for Pacific females aged 65-74 was 8 less.

Figure 20: Hierarchical tree of percentage of mortality records linked by sex, age, and ethnic group



† Numbers in brackets are the number of decedents (linked or unlinked) in each stratum, random rounded to an adjacent multiple of three.

‡ Percentage linked estimated by log-linear risk regression model, including two second order product terms: [sex] × [age group], and [age group] × [ethnic group]. The estimated percentages in bold are those that disagree by two more percent (absolute) from the actual.

Results from the stratified analyses suggested that there was generally little difference between 45-64 year olds and the 65-74 year olds in the characteristics influencing linkage to a census record, the above interaction of sex, age and ethnic group aside.

With consideration of prior information, and the desire for a parsimonious model, and the small absolute error between the first order interaction model and the actual numbers within strata, the log-linear risk model with just the two first order interactions was considered the 'best fit' for modeling the contribution of sex, age and ethnic group to linkage. This model was used as the benchmark for modeling in subsequent sections that attempt to determine the 'marginal' importance (not just statistical significance) of variables of interest (i.e. time period from census to death, cause of death, NZDep91 score, and NZSEI occupational class). The risk ratios for this model are shown in Table 25; the estimated percentage linked by strata are as shown previously in Figure 20, and are derived by multiplying the appropriate strata risk ratios by the intercept.

Table 25: Intercept and risk ratios for the final 'best-fit' log-linear risk model of mortality records linked to a census record, modeling for sex, age, and ethnic group

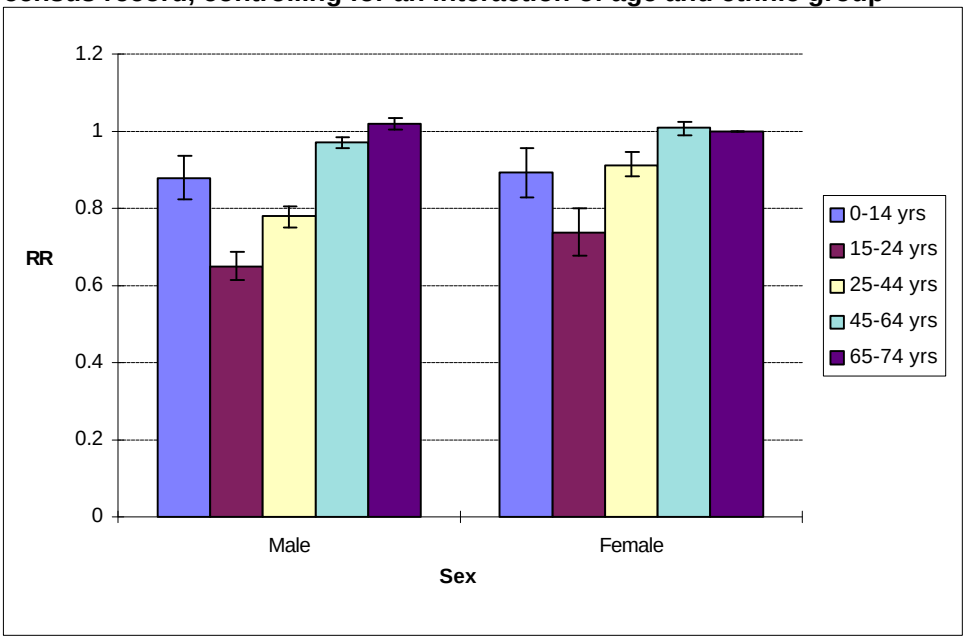
Variable or term			Risk ratio †	95% CI
Intercept			81.1%	(80.3-82.0%)
Sex × Age	Male	0-14	0.88	(0.82-0.94)
		15-24	0.65	(0.62-0.69)
		25-44	0.78	(0.75-0.80)
		45-64	0.97	(0.96-0.99)
		65-74	1.02	(1.00-1.03)
	Female	0-14	0.89	(0.83-0.96)
		15-24	0.74	(0.68-0.80)
		25-44	0.91	(0.88-0.94)
		45-64	1.01	(0.99-1.03)
		65-74‡	1.00	-
Age × Ethnic Group	0-14	Maori	0.82	(0.71-0.94)
		Pacific	0.99	(0.83-1.18)
		Rest‡	1.00	-
	15-24	Maori	0.94	(0.83-1.07)
		Pacific	0.85	(0.64-1.13)
		Rest‡	1.00	-
	25-44	Maori	0.87	(0.81-0.93)
		Pacific	0.79	(0.69-0.91)
		Rest‡	1.00	-
	45-64	Maori	0.85	(0.82-0.87)
		Pacific	0.72	(0.67-0.78)
		Rest‡	1.00	-
	65-74	Maori	0.75	(0.72-0.79)
		Pacific	0.73	(0.66-0.80)
		Rest‡	1.00	-

† The risk ratio is that compared to the referent group, except for the intercept which is the 'risk' of linkage for females aged 65-74 of non-Maori and non-Pacific ethnic group.
‡ Reference category.

Figure 21 and Figure 22 are graphical presentations of the risk ratios in Table 25. The two main qualitative observations from the table and figures are that:

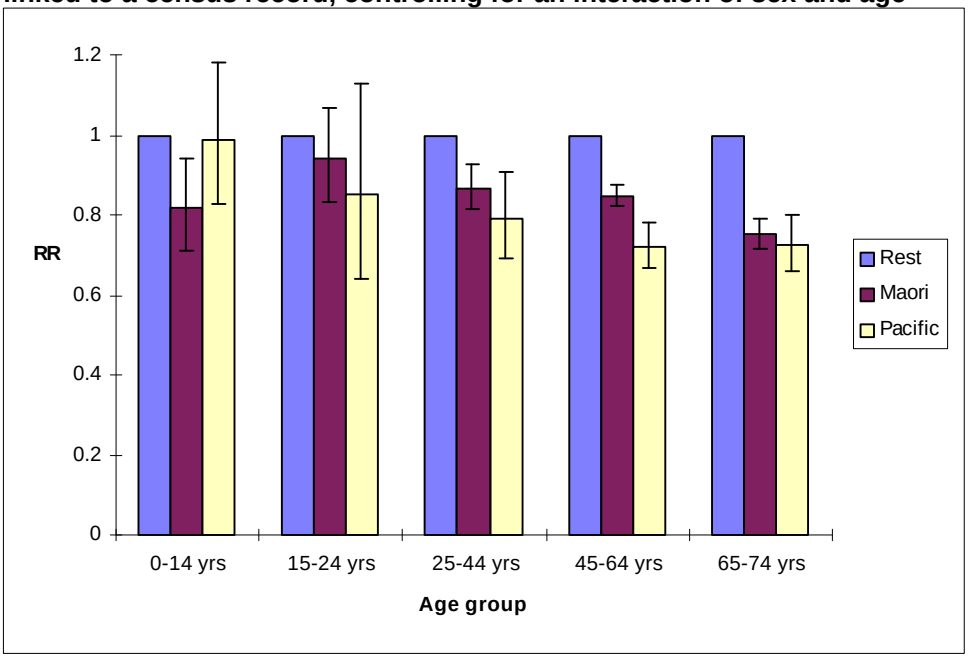
- the proportion of mortality records linked to a census record was comparable between males and females within age groups, except for 15-44 year old decedents where the proportions were lower for males compared to females
- the proportion of mortality records linked to a census record was less for Maori and Pacific decedents compared to the Rest, and this difference increased with increasing age.

Figure 21: Risk ratios compared to females aged 65-74 years demonstrating the interaction of sex and age on the estimated proportion of mortality records linked to a census record, controlling for an interaction of age and ethnic group



Error bars are the 95% confidence intervals for each risk ratio obtained from a log-linear model with just the two interaction products: [sex]×[age] and [age]×[ethnic group]. The reference category is females aged 65-74 years.

Figure 22: Risk ratios compared to non-Maori, non-Pacific demonstrating the interaction of age and ethnic group on the estimated proportion of mortality records linked to a census record, controlling for an interaction of sex and age



Error bars are the 95% confidence intervals for each risk ratio obtained from a log-linear model with just the two interaction products: [sex]×[age] and [age]×[ethnic group]. The reference category is the 'Rest' within each age group.

4.2.1.2 All cause mortality: time period between census and death

The relationship of time period between the census and death was investigated by considering:

- [time period] as the main effect of particular interest
- [sex × age] and [age × ethnic group] as the two 'main' effects to control for, based on the benchmark model from Section 4.2.1.1 above
- [time period] × [sex × age] and [time period] × [age × ethnic group] as the two interaction products of interest.

Backward elimination from this complete model found that the [time period] × [age × ethnic group] interaction product was statistically significant (p=0.0001 on Wald's Type III test), but the [time period] × [sex × age] was not (p=0.50). This finding was consistent with prior information from the univariate and stratified analyses that time period interacted with both age and ethnic group, but not sex. Thus the final model included [sex × age] and [time period] × [age × ethnic group].

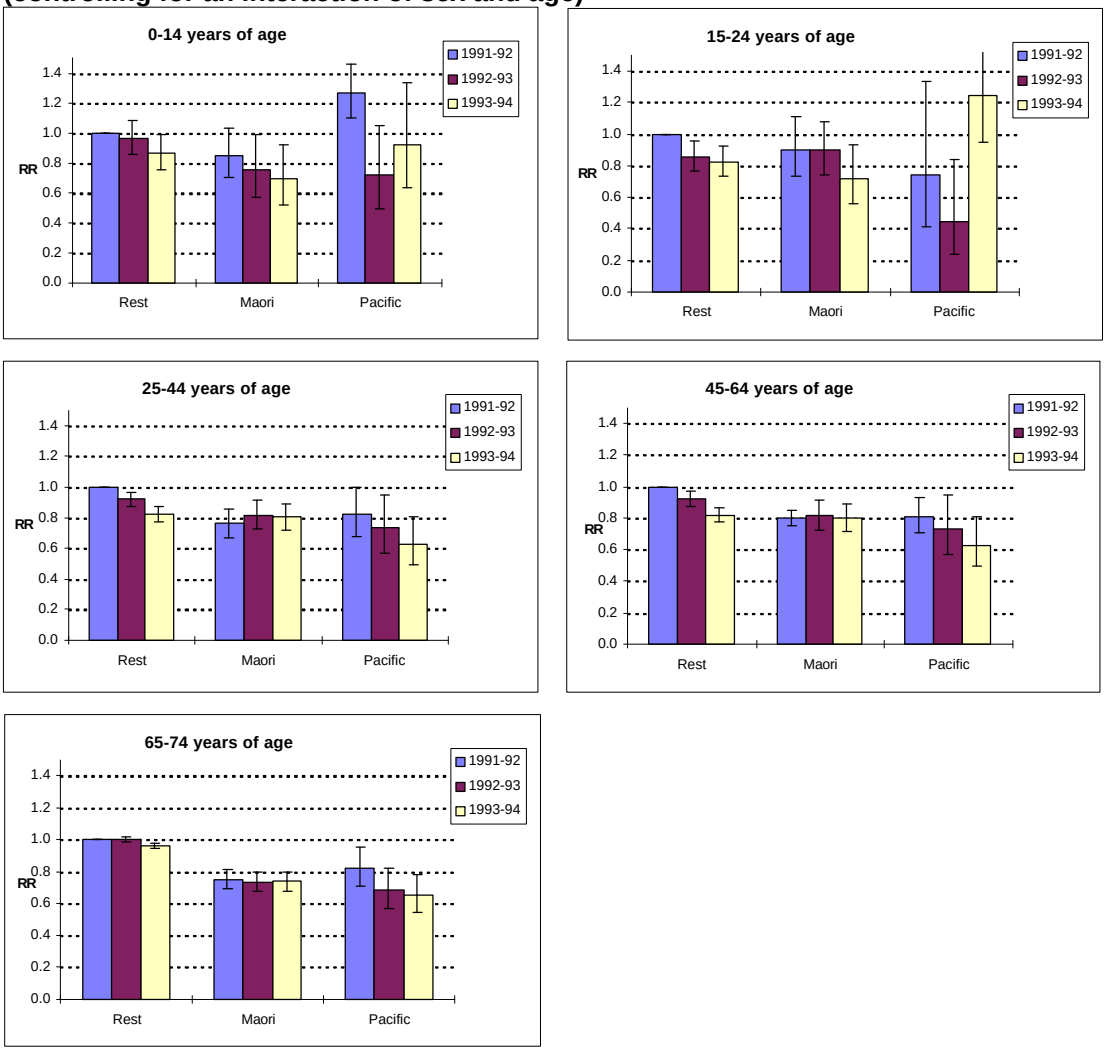
Results from this model are presented in Figure 23. The reference category for the risk ratios in each of the five age groups in Figure 23 is the linkage proportion for non-Maori and non-Pacific decedents (the 'Rest') dying in the first 12 months following the 1991 census. The main observations from Figure 23 are that:

- the proportion of non-Maori, non-Pacific (the 'Rest') decedents linked decreased over time following the census, more so for decedents aged less than 45 years than those aged 45 years and over
- there was little decrease over time in the proportion of Maori decedents aged 25 years and older linked
- there was a greater decrease over time in the proportion of Pacific decedents aged 25 years and older linked, compared to both Maori and the Rest.

Assuming that changing usual residence between the census and death is the main reason for decline in record linkage over time, then these results suggest that Maori are less mobile, and Pacific people more mobile, than the Rest - at least for people that are going to die in the next few years.

The risk ratios for Pacific decedents aged 0-24 are difficult to interpret due to small numbers (the confidence intervals are wide and mostly overlap for the three time periods).

Figure 23: Risk ratio for mortality records being linked to a census record for the interaction of time between the census and death, ethnic group, and age group (controlling for an interaction of sex and age)



The reference category for each sub-figure is the 'Rest' dying in the first year (1991-92) following the 1991 census. Error bars are the 95% confidence intervals for the risk ratios from a log-linear model.

4.2.2 NZDep91 small area deprivation

The objective of the analyses for bias by NZDep91 was to determine the bias by socio-economic factors in the record linkage, using the NZDep91 score:

- as a direct measure of the socio-economic construct 'small area deprivation'

- **and,** as a proxy measure for other socio-economic factors that are likely to be correlated with NZDep91, but for which there was no measure on the mortality data.

4.2.2.1 Bias by NZDep91, controlling for sex, age and ethnic group

This analysis of bias by NZDep91 score, controlling for sex, age, and ethnic group, is the most important of the analyses of bias in the record linkage, being a specific analysis of bias by socio-economic factors including 89.4% of all the mortality records (n=36,927).

NZDep91 was specified as either a decile or quintile categorical variable. The initial model for backward elimination included five effects or products:

- [NZDep91 quintile] as the main effect of interest
- [sex × age] and [age × ethnic group] as the main effects to control for
- [NZDep91 quintile] × [sex × age], and [NZDep91 quintile] × [age × ethnic group] as the two interaction products of interest.

Neither of the latter two interaction products were statistically significant on Wald's Type III test (p=0.66 and p=0.28 respectively). Thus the final model was simply the first three 'main effects' in the list above, suggesting that the bias in record linkage by small area deprivation did not vary by strata of sex, age and ethnic group. This 'uniform' NZDep91 bias by strata is shown in Table 26. The risk ratios range from 0.98 to 0.96 only for decile 2 to decile 9, then drops off to 0.92 for the most deprived decile compared to the least deprived decile.

The final model fits well with prior information. Whilst association of NZDep91 quintile with linkage was confounded by ethnic group, there was no evidence on stratified analyses that NZDep91 score interacted with sex, age or ethnic group.

Table 26: Risk ratios by NZDep91 decile for the percentage of mortality records linked, controlling for sex, age, and ethnic group in a log-linear model[‡]

Variable		Risk ratio	95% CI
NZDep91 decile	1[†]	1.00	
	2	0.98	(0.96-1.01)
	3	0.98	(0.96-1.01)
	4	0.98	(0.95-1.00)
	5	0.98	(0.96-1.00)
	6	0.97	(0.95-1.00)
	7	0.96	(0.94-0.98)
	8	0.96	(0.94-0.98)
	9	0.96	(0.94-0.98)
	10	0.92	(0.90-0.94)

[†] Reference category, the least deprived (whereas decile 10 is the most deprived).
[‡] The log-linear model included [sex ∗ age group], [age group ∗ ethnic group], and [NZDep91I].

Using the NZDep91 scores as a proxy for socio-economic factors generally, this result suggests that **there is little bias in the record linkage by socio-economic factors controlling for sex, age and ethnic group**. This has important implications for the cohort study. First, controlling for sex, age and ethnic group, there should be little *follow-up bias* by socio-economic factors in the cohort study. Second, any small follow-up bias by socio-economic factors does not vary discernibly by strata of sex, age, and ethnic group.

Given the importance to the research of this analysis of the NZDep91 scores, its robustness was examined further by using an alternative model selection approach, and considering the possible bias from not including the 10.6% of the mortality records that had no NZDep91 score (n=4,383).

An alternative log-linear model was developed as per the general strategy outlined on page 67, but including just the six first order interaction products formed by considering sex, age, ethnic group, and NZDep91 quintile as four separate main effects. (That is, the previous finding that sex, age, and ethnic group contributed to linkage as two interactions of [sex × age] and [age × ethnic group] was ignored.) The final model selected by backwards elimination was the same as that above: [sex × age], [age × ethnic group], and [NZDep91], suggesting the model was robust. The interaction product of [ethnic group × NZDep91] approached statistical significance

(p=0.05 to 0.10 depending on whether the NZDep91 score was specified as a decile or quintile, and whether the reference category was most deprived or least deprived), but was not considered further for inclusion due to the suspected artefact in this interaction product as discussed on page 111.

Could the final model above have changed if a NZDep91 score had been assigned to the 10.6% of mortality records for which there was no meshblock? Probably not. First, it was unlikely that a model for 89.4% of the mortality records will substantially differ from that for all the mortality records. Second and more specifically, for the extra 10.6% of mortality records to affect the overall model, the distribution of linkage rates by socio-economic factors would probably have to vary substantially compared to that for the 89.4% already modeled. Inspection of the percentage linked by NZSEI occupational class for mortality records with no NZDep91 score assigned found a similar relationship to that for mortality records with a NZDep91 score assigned, thus discounting this possibility.

4.2.2.2 Bias by NZDep91, controlling for time period, sex, age and ethnic group

The objective of this analysis was to determine whether the small socio-economic bias, as measured by NZDep91, varied by year of follow-up. The reason for undertaking this analysis was that the cohort study will include analyses of the association of socio-economic factors with death for different follow-up periods (eg to investigate possible health selection effects). We want to know whether differences (if any) in the association of socio-economic factors with death over time are, in part at least, due to bias incurred in the record linkage process. As the distribution of variables for deaths (mainly NZDep91, but also sex, age and ethnic group) will not vary much over the short three year follow-up period, confounding is not of interest. Instead, interaction of time period and NZDep91 score is of interest, indicating a bias by NZDep91 score that varies by year of follow-up.

Prior information that was available for consideration in this analysis included the following:

- stratified analyses of time period between census and death by NZDep91 quintile (and NZSEI occupational class) suggested no interaction (page 98)
- the 'best-fit' log-linear model for the effect of time period on linkage, controlling for sex, age, and ethnic group, consisted of [sex × age] and [age × ethnic group × time period]
- the 'best-fit' log-linear model for the effect of NZDep91 score on linkage, controlling for sex, age, and ethnic group, consisted of [sex × age], [age × ethnic group], and [NZDep91 decile].

Given the main effects in the latter two bullet points, a theoretically reasonable initial model was [sex × age], [age × ethnic group × time period] and [NZDep91] as the main effects, and [age × ethnic group × time period] × [NZDep91 quintile] as the interaction product of interest. However, and not surprisingly, a log-linear model including these main effects and interaction product failed to converge. (It would also have been unwieldy to interpret.)

A simplified model was therefore specified. First, model selection was conducted separately by sex. Second, [age × ethnic group × time period] and [NZDep91] were retained as the main effects. But the interaction product of interest was simplified to just [time period (years)] × [NZDep91 quintile]. These variables again failed to converge in a log-linear model (invalid values estimated), but a logistic model could be fitted. For both sexes, the two main effects were statistically significant on Wald's Type II test (p=0.0001 for both for males, p=0.0001 and p=0.0282 for females respectively), but the interaction product was not (p=0.42 for males, p=0.27 for females). Whilst the logistic model will give different results to a log-linear model, it seemed reasonable to conclude that the interaction product was not important.

To test the robustness of the above finding, an alternative log-linear model was fitted separately by sex, treating age, ethnic group, time period and NZDep91 score as four separate main effects, and specifying their six first order interaction products. The interaction product of interest, [NZDep91 quintile] × [time period (years)], was not

statistically significant on Wald's Type III test for males (p=0.38) nor females (p=0.20).

The conclusion was that the bias in record linkage by NZDep91 probably did not vary by time between the census and death. The implication for the cohort study is that if any variation is observed in the relative association of socio-economic factors with mortality over time, it would probably not be due to bias in the record linkage.

4.2.2.3 Bias by NZDep91 by cause of death, controlling for sex, age and ethnic group

Cancer

Cancer deaths were limited to those for 25-74 year olds due to insufficient numbers for ethnic and sex strata in younger ages. Considering just [age], [sex] and [ethnic group], the interaction product of [sex] × [age group] was statistically significant on log-linear modeling, but not [age] × [ethnic group]. Thus, the inter-relationships between these three demographic variables were different for cancer deaths compared to all deaths. Considering [sex × age group], [ethnic group], and [NZDep91] as the main effects, there were no significant interactions between [NZDep91] and the two other main effects. Thus the bias by small area deprivation in the record linkage for cancer deaths was independent of age, sex and ethnic group. The risk ratios are shown in Table 27, and suggest little, if any (except for decile 10), decline in the percentage of cancer deaths linked by NZDep91 decile, controlling for sex, age and ethnic group.

Table 27: Risk ratios by NZDep91 decile for the percentage of mortality records linked by cause of death (cancer, ischaemic heart disease, and unintentional injury), controlling for sex, age, and ethnic group

	Cancer (n=12,389; 25-74 yrs)		IHD (n=8,999; 25-74 yrs)		Unintentional injury (n=2241; 0-74 yrs)	
NZDep91 decile	Risk ratio	95% CI	Risk ratio	95% CI	Risk ratio	95% CI
1[†]	1.00		1.00		1.00	
2	1.00	(0.96-1.03)	0.95	(0.91-0.99)	1.00	(0.89-1.13)
3	1.00	(0.96-1.03)	0.99	(0.95-1.03)	0.85	(0.73-0.98)
4	0.98	(0.95-1.02)	0.97	(0.93-1.01)	0.91	(0.80-1.03)
5	0.97	(0.94-1.01)	0.96	(0.92-1.00)	0.92	(0.81-1.05)
6	1.00	(0.97-1.03)	0.93	(0.89-0.97)	0.98	(0.88-1.10)
7	0.98	(0.95-1.01)	0.95	(0.91-0.99)	0.90	(0.79-1.02)
8	0.98	(0.95-1.01)	0.93	(0.89-0.97)	0.90	(0.79-1.01)
9	0.97	(0.94-1.00)	0.94	(0.90-0.98)	0.86	(0.76-0.98)
10	0.94	(0.90-0.98)	0.90	(0.87-0.94)	0.86	(0.76-0.98)

[†] Reference category

Ischaemic heart disease

Ischaemic heart disease (IHD) deaths were also limited to 25-74 year olds. Modeling just sex, age and ethnic group, only [age group] × [ethnic group] was a significant interaction product. Considering [sex], [age group × ethnic group], and [NZDep91] as the main effects, neither first order interaction involving [NZDep91], the main effect of interest, was statistically significant. Risk ratios for NZDep91 deciles are shown in Table 27, and suggest a decline in the percentage of IHD deaths linked by NZDep91 decile similar to that for all deaths, controlling for sex, age and ethnic group.

Cardiovascular disease (other than ischaemic heart disease)

Log-linear modeling of sex, age, ethnic group and NZDep91 for other cardiovascular deaths (n=5306) found a statistically significant interaction of [NZDep91] and [age group] (p=0.01 on Walds Type III test) when main effects were considered to be [sex × ethnic group] and [age group] (based on modeling for just sex, age and ethnic group), and the main effect of interest [NZDep91]. This interaction term was also significant if sex, age, ethnic group and NZDep91 were all just treated as separate main effects. However, closer inspection of the interaction product demonstrated that it was only significant due to an erratic distribution of linkage success by small area deprivation for 25-44 year olds: for the two older age groups there was a mild decline (5-10%) in linkage with increasing small area deprivation. (Results not presented.)

Injury

All ages of unintentional injury death were included. Modeling just sex, age and ethnic group, none of the possible first order interaction products were statistically significant. Considering [sex], [age group], [ethnic group], and [NZDep91] as the main effects, none of the possible first order interaction involving [NZDep91], the main effect of interest, were statistically significant. Risk ratios for NZDep91 deciles are shown in Table 27, and suggest a moderate (albeit erratic) decline in the percentage of unintentional injury deaths linked by NZDep91 decile, controlling for sex, age and ethnic group.

Suicide

Numbers were fewer for suicide deaths (n=1435) than other causes. Modeling found an erratic distribution of mortality records linked by small area deprivation controlling for sex, age and ethnic group, and 95% confidence intervals for the risk ratios all included 1.0. There was no apparent decrease in increase in the percentage of suicide deaths linked with increasing small area deprivation. Thus, as best can be judged in this project, there was no obvious bias by small area deprivation in the chance of suicide deaths being linked.

Other causes of death

There were too few deaths from infection (n=1013) and interpersonal violence (n=186) to model, and it made little sense to consider the remaining deaths (n=4180) as a common group for modeling.

4.2.3 Whether occupation was recorded on the death registration form

The objective of the analyses for bias by whether an occupation was recorded on the BDM28 was to determine the bias by employment status in the record linkage, using whether or not occupation was recorded as a proxy measure for employment status among decedents age 15 years and older. As such, it should be treated with considerable caution. The interpretation also varies by sex, so all modeling was conducted separately by sex.

Whether occupation was recorded on the BDM28 was specified as a binary variable (occupation), with the reference category being having an occupation. For males, 20,548 of the 24,724 aged 15 years and older had an occupation recorded on the BDM28 (83.1%) . However, for females only 2,989 of the 15,755 aged 15 years and older had an occupation recorded (19.0%).

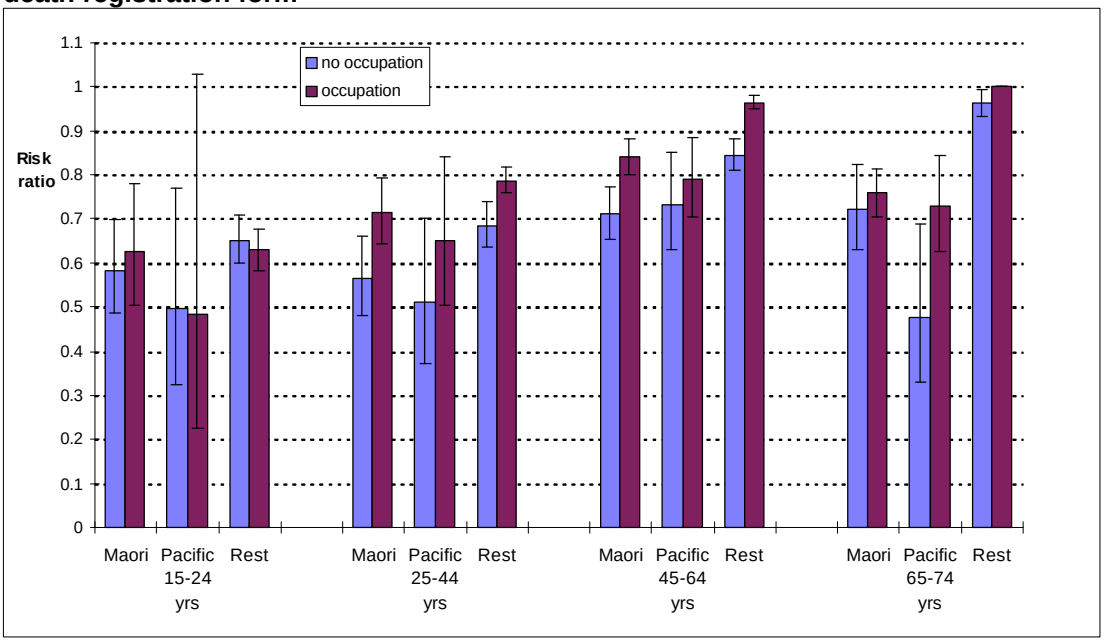
Using the same rationale as in previous sections, and all deaths, the initial model for backward elimination included:

- [occupation] as the main effect of interest
- [age × ethnic group] as the main effect to control for (from previous modeling of just age and ethnic group for all deaths)
- [occupation] × [age × ethnic group] as the interaction product of interest.

The interaction term was statistically significant at the five percent level on Wald's Type III test for both males (p=0.0022) and females (p=0.045). The results for males are shown in Figure 24 as risk ratios for linkage by strata of age, ethnic group and occupation, with the reference category being non-Maori, non-Pacific aged 65-74 with a stated occupation. Perhaps the main observation from Figure 24 is that males aged 25-64 with an occupation are more likely to be linked to a census record than similar aged males with no occupation. Indeed, a log-linear model with age, ethnic group and occupation specified as three separate main effects, and their three first order interaction products, only the interaction product of [age] × [occupation] was statistically significant. In this simplified model, males aged 25-44 without an occupation had a risk ratio of 0.85 (95% confidence interval 0.79-0.92) compared to males of the same age with an occupation of being linked to a census record, and the equivalent risk ratio for males aged 45-64 was 0.87 (0.84-0.91).)

The second, but minor, observation from Figure 24 was that male Pacific decedents aged 45-64 years and 65-74 years again appeared to deviate from the pattern for the other two ethnic groups. That is, the difference between those with and without an occupation was less for 45-64 year olds, but greater for 65-74 year olds. However, the numbers are small, as reflected by the wide confidence intervals.

Figure 24: Risk of male mortality records being linked to a census record for the interaction of age, ethnic group, and whether an occupation was recorded on the death registration form



The reference category is the 'Rest' aged 65-74.
Error bars are the 95% confidence intervals for the risk ratios from a log-linear model.

If not having an occupation recorded on the BDM28 can be considered as a proxy for unemployment, then the implication of this result for the cohort study is that any association of employment status with death for 25-64 year old males (at least) will be biased, perhaps by about 15%. However, as mentioned above, this result needs to be treated cautiously due to the many assumptions entailed.

As for males, the interaction of [occupation] × [age × ethnic group] was statistically significant for females at the five percent level, although only just (p=0.045). Inspection of a histogram for females, equivalent to that for males in Figure 24, was not informative (and is not presented here) - there was no apparent pattern. Of note, the main observation for 25-64 year olds males of a greater difference in linkage between those with and without an occupation compared to other age groups was not reproduced. Confidence intervals for individual risk ratios were also often wide. Given the lack of any identifiable pattern, the marginal statistical significance of the interaction, and the even more tenuous use of recorded occupation as a proxy for employment status in females compared to males, the interaction term was rejected.

The final model for females was therefore just [age × ethnic group] and [occupation]. The risk ratio for females with no occupation compared to females with an occupation was 0.99 (95% confidence interval 0.97-1.01) - but it must be treated with great caution as a proxy measure of employment status.

4.2.4 NZSEI occupational class

The objective of the analyses for bias by NZSEI occupational class was analogous to that by NZDep91. That is, to determine the bias by socio-economic factors in the record linkage, using the NZSEI occupational class:

- as a direct measure of the socio-economic construct 'occupational class'
- **and,** as a proxy measure for other socio-economic factors that are likely to be correlated with occupational class, but for which there was no measure on the mortality data.

Analyses of bias by NZSEI occupational class for specific causes of death were not conducted. Problems with such analyses would include relatively small numbers by sex (particularly females) for specific causes of death, and consequent difficulty determining if there was any meaningful variation by cause of death. Instead, the previous NZDep91 analyses of bias by cause of death should be referred to, being for substantially greater numbers of deaths, and for sexes combined.

For males, 13,701 of the 16,077 decedents aged 25-74 years on census night, and dying in the second and third year of follow-up, had an occupation recorded on the BDM28 (84.2%). However, for females only 2,059 of the 10,496 similarly restricted decedents had an occupation recorded (19.6%).

The analyses are conducted separately by sex, and attempted for decedents aged 25-74 years on census night and dying in the second and third year of follow-up.

4.2.4.1 Males, controlling for age and ethnic group

The main effects considered in this analysis, based on previous analyses, were [age group × ethnic group] and [NZSEI]. Thus the interaction of interest was [age group × ethnic group] × [NZSEI]. Numerous exploratory models pointed to the same conclusion: NZSEI occupational class did not interact with age and ethnic group in the modeling of mortality records linked, but had an independent effect. The risk ratios for linkage to a census record by NZSEI occupational classes are shown in Table 28. There was no obvious drop in linkage rates between occupational classes 1 to 5, but a 6% relative drop from occupational class 5 to 6. The lower linkage rate for farmers is presumably a function of their rural status, and consequent lower probability of a meshblock code.

Table 28: Risk ratios by NZSEI occupational class for the percentage of mortality records linked by sex for 25-74 year olds, controlling for age and ethnic group

NZSEI occupational class	Males (n=13,701) ‡		Females (n=2,059) ‡	
	Risk ratio	95% CI	Risk ratio	95% CI
1	1.01	(0.97-1.04)	0.97	(0.87-1.08)
2	1.02	(0.99-1.05)	0.97	(0.91-1.03)
3	1.00	(0.97-1.02)	0.98	(0.92-1.05)
4†	1.00	-	1.00	-
5	1.00	(0.97-1.02)	0.98	(0.91-1.05)
6	0.94	(0.91-0.98)	0.91	(0.84-0.99)
Farmers	0.91	(0.88-0.94)	0.86	(0.74-0.99)

† Reference category

‡ For both sexes, the risk ratios are from a log-linear model with just the interaction product [age group * ethnic group] and the main effect [NZSEI].

4.2.4.2 Females, controlling for age and ethnic group

Exploratory analyses suggested that there was no interaction between NZSEI occupational class and age or ethnic group. Risk ratios for the final model are shown in Table 28. As for males, there was little difference across occupational classes 1 to 5, but then a 7% relative drop from occupational class 5 to 6.

4.2.5 NZDep91 small area deprivation and NZSEI occupational class considered simultaneously

The objective of this analysis was to determine the independent bias for NZDep91 small area deprivation and NZSEI occupational class. Analyses were conducted separately by sex for the for 12,249 male and 1,884 female decedents aged 25-74 years with both a NZDep91 score and NZSEI occupational class. Covariates were age and ethnic group, but no attempt was made to also control for time period.

4.2.5.1 Males, controlling for age and ethnic group

The interaction of small area deprivation and occupational class in the prediction of linkage success was assessed using various groupings of NZDep91 (quintiles; three groups formed by combining deciles 1-3, 4-7, and 8-10) and NZSEI (six classes and farmers; six classes ignoring farmers; three groups formed by combining classes 1-2, 3-4, and 5-6). These groupings of NZDep91 and NZSEI were then cross-classified to form an array of dummy variables. For example, cross-classifying NZDep91 quintile and the three groups of combined NZSEI occupational classes resulted in $5 \times 3 = 15$ possible cross-classified categories, which reduced to 14 dummy variables. None of these cross-classified 'interaction' variables of NZDep91 and NZSEI had a statistically significant effect, over and above the main effects of NZDep91 and NZSEI ($p>0.16$ on Wald's Type III test in all instances). This lack of any interaction suggested that the bias in the record linkage for males from NZDep91 and NZSEI occupational class were independent of each other. Risk ratios by NZDep91 deciles and NZSEI occupational class as separate main effects are shown in Table 29. There was a significant decrease in record linkage success for decedents in the two most deprived deciles of small areas (risk ratio = 0.95) and for decedents in occupational class 6 (risk ratio = 0.95). Otherwise, there is little substantive difference by socio-economic status in the record linkage.

Table 29: Risk ratios of NZDep91 decile and NZSEI occupational class for linkage to a census record by sex (n=12,249 male decedents and n=1,884 female decedents) [†]

		Males		Females	
		Risk ratio	95% CI	Risk ratio	95% CI
NZDep91 Decile	1[‡]	1.00		1.00	
	2	1.00	(0.97-1.04)	1.09	(1.00-1.19)
	3	1.00	(0.96-1.03)	1.03	(0.94-1.13)
	4	0.99	(0.95-1.03)	1.06	(0.96-1.17)
	5	0.98	(0.95-1.02)	1.02	(0.92-1.12)
	6	0.99	(0.95-1.03)	1.03	(0.94-1.13)
	7	0.99	(0.95-1.03)	0.95	(0.86-1.05)
	8	0.97	(0.93-1.01)	1.01	(0.92-1.11)
	9	0.95	(0.92-0.99)	0.99	(0.90-1.09)
	10	0.95	(0.92-0.99)	0.99	(0.89-1.11)
Occupational class	1	1.00	(0.97-1.03)	0.96	(0.87-1.07)
	2	1.01	(0.98-1.04)	0.96	(0.90-1.03)
	3	1.00	(0.97-1.02)	0.97	(0.91-1.04)
	4[†]	1.00		1.00	
	5	0.99	(0.97-1.02)	0.97	(0.91-1.04)
	6	0.95	(0.92-0.99)	0.92	(0.85-1.00)
	Farmers	0.93	(0.90-0.97)	0.99	(0.87-1.13)

[†] Restricted to decedents aged 25-74 on census night dying in the second and third year of follow-up, and with non-missing data for both NZDep91 and NZSEI. The risk ratios are from a log-linear regression model including [NZDep91], [NZSEI], and [age group * ethnic group] as independent variables.
[‡] Reference category.

The risk ratios in Table 29 are not directly comparable with results for NZDep91 and NZSEI alone (Table 26, page 128, and Table 28, page 137, respectively). The results in Table 29 are for the restricted sample of decedents aged 25-74 on census night, dying in the second and third year of follow-up, and with non-missing data for both NZDep91 and NZSEI. Using this restricted sample, and modeling NZDep91 and NZSEI separately, the risk ratios were not substantively different from those in Table 29, with the possible exception of a small reduction to the null for NZDep91 decile risk ratios when controlling for occupational class. It is tempting to conclude, therefore, that NZDep91 and NZSEI each have independent effects on linkage success that are little confounded by the other. However, as the bias in linkage by either NZDep91 or NZSEI is modest (if not null for most socio-economic strata), such a conclusion may be an over-interpretation. A more reasonable conclusion is that there was little bias in the record linkage by socio-economic status, *except* for the lowest socio-economic groups. Considering the lowest socio-economic groups, a male decedent from the two most deprived deciles of small areas and in occupational class 6 had a 10% reduced chance of being linked to a census record ((1-[0.95×0.95]) / 1) compared to a male

decedent from the least deprived decile of small areas and in occupational class 1. But for all other combinations of small area deprivation and occupational class for males, the bias is at most 5% and usually much less.

4.2.5.2 Females, controlling for age and ethnic group

As with males, the interaction [NZSEI] × [NZDep91] was not statistically significant. Results for NZDep91 deciles and NZSEI occupational class as separate main effects predicting linkage success are shown in Table 29. All confidence intervals include 1.0, but the trends are consistent with those for males. Most importantly though, the results should be treated with caution as only 17.9% of female decedents aged 25-74 on census night and dying in the second and third year of follow-up had both a NZDep91 and NZSEI score.

Chapter 5: Conclusion

The aim of this research was:

- to determine the feasibility of anonymously linking census and mortality records using probabilistic record linkage software (Automatch®).

We believe we have demonstrated that anonymous record linkage is feasible, and that Automatch® works reasonably well.

The objectives of this report were:

- to determine the percentage of mortality records that could be linked to a census record.
- to determine the accuracy of the linkage as measured by the positive predictive value (i.e. the percentage of all links accepted that were estimated to be correct links of the same individual's census and mortality record).
- to analyse the variation by demographic and socio-economic factors, between linked and unlinked mortality records, and hence estimate the bias in the record linkage (i.e. the relative difference in probability of linkage for a high compared to low socio-economic individual).

And with a target:

- at least 70 percent of mortality records for deaths six to 18 months following the census should be successfully linked to a census record

or

- 60 to 70 percent of deaths six to 18 months following the census should be successfully linked to a census record, with little apparent bias by socio-economic factors between linked and unlinked mortality records.

We managed to link 76.6% of mortality records to a census record *for deaths in the full three-year follow-up period*, with a positive predictive value of greater than 95%. Thus we successfully met our target, and achieved high accuracy in the record linkage.

Regarding bias in the record linkage, there was notable variation in the linkage by demographic characteristics. However, within demographic strata there was relatively little evidence of a bias by socio-economic status, except for the lowest socio-economic groups. Thus, there will be a modest small bias in the follow-up of the 1991 census cohort by socio-economic status. This has to be allowed for in the subsequent cohort analyses.

The aim of the NZCMS is not just to conduct record linkage – the record linkage is a means to an end. Given that the record linkage we attempted pushed the technical feasibility of anonymous and probabilistic record linkage, we have produced this report to thoroughly document that process. Focus must now shift to the substantive cohort analyses.

Appendix: NZSEI Occupational Class

The New Zealand Socio-Economic Index (NZSEI) assigns a standardised score between 10 and 90 for each of the 97 minor occupations in the 1990 New Zealand Standard Classifications of Occupations (NZSCO90).[30] Davis et al (1997) proposed that for categorical analyses, the scores could be divided as given in Table 30. Davis et al stated that their division into six occupational classes was a starting point: we have chosen to modify the classification as shown in Table 30.

Table 30: Alternative classifications of 'occupational class' from NZSEI scores

	Range of NZSEI scores (% of 20-69 year old population 1991 census †)		% of 15-64 year old males by Elley-Irving
Occupational class	Davis et al, 1997	Modification	Class, 1986 census ‡
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 A of Davis et al (1997), giving similar but not identical results to that shown in Table 3.8 of Davis et al.[30]
‡ Taken from Pearce et al, 1991.[7]
# 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.

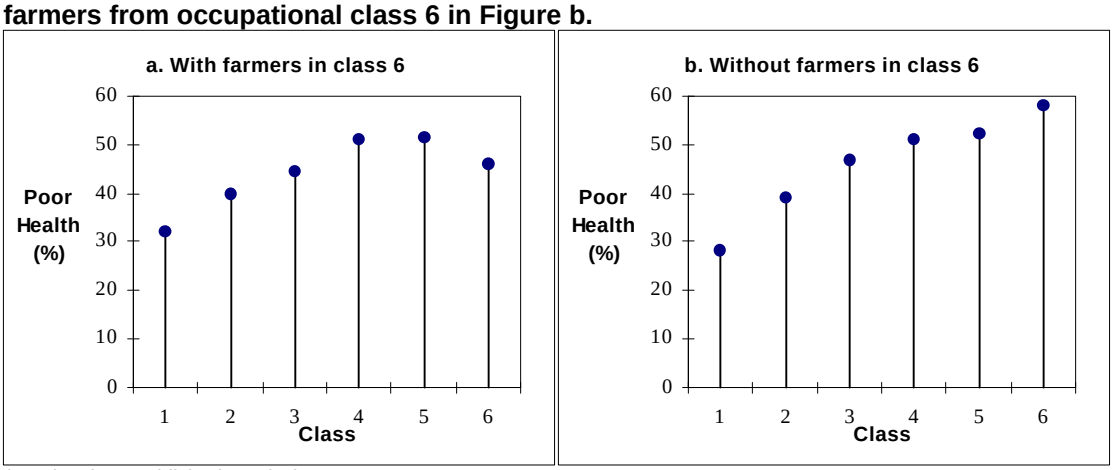
There were three reasons why we preferred the modified NZSEI occupational class classification. First, close inspection of NZSCO90 codes allocated by Davis et al to occupational class 2 disclosed a bimodal distribution of NZSEI scores - NZSEI scores were either between 60 and 65, or between 70 and 75. 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 socio-economic status 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) that the former occupations with a NZSEI score between 60 and 65.

Second, the percentage distribution for the modified classification (excluding farmers) is probably better that that proposed by Davis et al. The distribution is more symmetric, 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 more closer to that for the Elley-Irving scale.

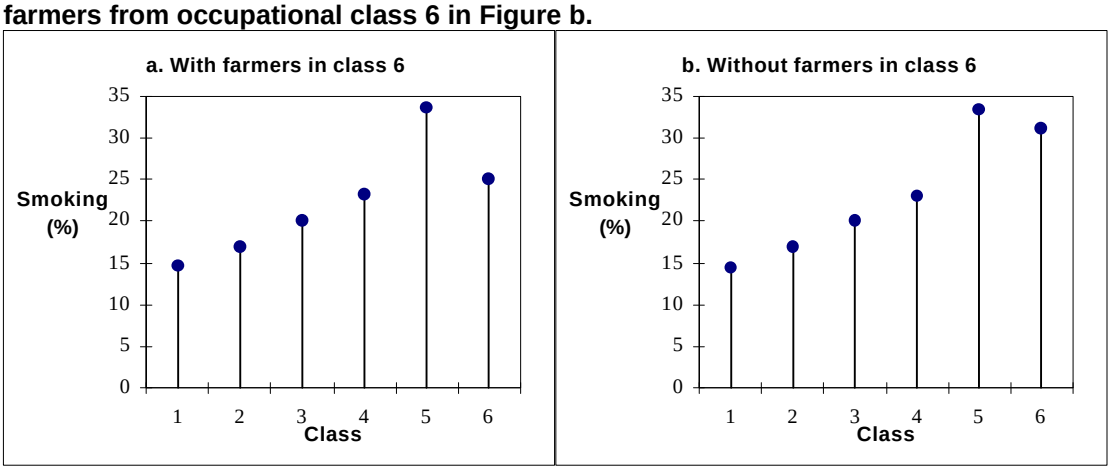
Third, there are problems with ranking farmer's socio-economic status that probably argue for their removal from the ordinal ranking of occupational classes to be considered as a separate 'special' group. The NZSCO90 classification has just two minor codes for farmers: NZSCO90 code 611(market farmers 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 socio-economic status.[30] Moreover, farm owners are also self-employed, a group known to have a low declared income compared to similar status occupations in New Zealand.[33] Occupational class indices used in Europe commonly separate farmers into a separate occupational class (eg [34]). As generated by the path model used to develop the NZSEI score,[30] the two farming occupation codes both fell within occupational class 6 and, furthermore, farmers comprised more than half of Davis et als' proposed occupational class 6. 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 25 and Figure 26 demonstrate this effect.

Figure 25: 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.



(Previously unpublished results.)

Figure 26: 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.



(Previously unpublished results.)

A similar problem probably also exists for the Armed Forces,[30] 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 NZSEI scores. For example, farmers may be successfully separated and differentiated along a socio-economic status continuum.

References

1. Last J. *A dictionary of epidemiology*. 3 ed. New York: Oxford University Press, 1995.

2. Pearce N, Davis P, Smith A, Foster F. Mortality and social class in New Zealand. I: overall male mortality. *NZ Med J* 1983;**96**:281-285.

3. Pearce N, Davis P, Smith A, Foster F. Mortality and social class in New Zealand. II: male mortality by major disease groupings. *NZ Med J* 1983;**96**:711-716.

4. Pearce N, Davis P, Smith A, Foster F. Mortality and social class in New Zealand. III: male mortality by ethnic group. *NZ Med J* 1984;**97**:31-35.

5. Pearce N, Davis P, Smith A, Foster F. Social class, ethnic group, and male mortality in New Zealand, 1974-8. *J Epidemiol Community Health* 1985;**39**:9-14.

6. Pearce N, Howard J. Occupation, social class and male cancer mortality in New Zealand, 1974-78. *Int J Epidemiol* 1986;**15**:456-462.

7. Pearce N, Marshall S, Borman B. Undiminished social class mortality differences in New Zealand men. *NZ Med J* 1991;**104**:153-156.

8. Pearce N, Pomare E, Marshall S, Borman B. Mortality and social class in Maori and nonMaori New Zealand men: changes between 1975-7 and 1985-7. *NZ Med J* 1993;**106**:193-196.

9. Pearce N, Bethwaite P. Social class and male cancer mortality in New Zealand, 1984-7. *NZ Med J* 1997;**110**:200-202.

10. Kawachi I, Marshall S, Pearce N. Social class inequalities in the decline of coronary heart disease among New Zealand men, 1975-1977 to 1985-1987. *Int J Epidemiol* 1991;**20**:393-398.

11. Marshall S, Kawachi I, Pearce N, Borman B. Social class differences in mortality from diseases amenable to medical intervention in New Zealand. *Int J Epidemiol* 1993;**22**(2):255-261.

12. Jackson G, Kelsall L, Parr A, Papa D. Socio-economic inequalities in health care. Auckland: North Health, a Division of the Health Funding Authority, 1998.

13. Crampton P, Salmond C, Sutton F. NZDep91: a new index of deprivation. *Social Policy Journal of New Zealand* 1997;**9**:186-193.

14. Salmond C, Crampton P, Sutton F. NZDep91: A New Zealand index of deprivation. *Aust NZ J Public Health* 1998;**22**:835-837.

15. Fox A, Goldblatt P. Longitudinal Study 1971-1975: Socio-demographic mortality differences. London: Her Majesty's Stationary Office: Office of Population Censuses and Surveys, 1982.

16. Faggiano F, Lemma P, Costa G, Gnavi R, Pagnanelli F. Cancer mortality by educational level in Italy. *Cancer Causes and Control* 1995;**6**:311-320.

17. Rogot E, Sorlie P, Johnson N. Probabilistic methods in matching census samples to the national death index. *J Chron Dis* 1986;**39**:719-734.

18. Calle E, Terrell D. Utility of the National Death Index for ascertainment of mortality amog Cancer Prevention Study II Participants. *Am J Epidemiol* 1993;**137**(2):235-241.

19. Vagero D, Lundberg O. Health inequalities in Britiain and Sweden. *Lancet* 1989;**July 1**:35-36.

20. Westerling R, Gullberg A, Rosen M. Socioeconomic differences in 'avoidable' mortality in Sweden 1986-1990. *Int J Epidemiol* 1996;**25**:560-567.

21. Leon D, Lithell H, Vagero D, Koupilova I, Mohsen R, Berglund L, et al. Reduced fetal growth rate and increased risk of death from ischaemic health disease: cohort study of 15 000 Swedish men and women born 1915-29. *BMJ* 1998;**317**:241-245.

22. Newcombe H. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press, 1988.

23. Jaro M. Probabilistic linkage of large public health data files. *Stat Med* 1995;**14**:491-498.

24. Baldwin J, Acheson E, Graham W. *Textbook of medical record linkage*. Oxford: Oxford University Press, 1987.

25. Pomare E, Keefe-Ormsby V, Ormsby C, Pearce N, Reid P, Robson B, et al. *Hauora: Maori Standards of Health III*. Wellington: Eru Pomare Maori Health Research Centre, 1995.

26. Newcombe H. Age-related bias in probabilistic death searches due to neglect of the "Prior Likelihoods". *Computers and Biomedial Research* 1995;**28**:87-99.

27. MatchWare Technologies I. Automatch Generalised Record Linkage System, Version 4.2: User's Manual. Kennebunk, Maine: MatchWare Technologies, Inc, 1998.

28. Newcombe H, Smith M, Howe G, Mingay J, Strugnell A, Abbatt J. Reliability of computerized versus manual death searches in a study of the health of Eldarado uranium workers. *Comput Biol Med* 1983;**13**:157-169.

29. Roos LJ, Wajda A, Nicol J. The art and science of record linkage: methods that work with few identifiers. *Computers in Biology and Medicine* 1986;**16**:45-57.

30. Davis P, McLeod K, Ransom M, Ongley P. The New Zealand Socioeconomic Index of Occupational Status (NZSEI). Wellington: Statistics New Zealand, 1997.

31. Murray J, Lopez A. *The Global Burden of Disease: A comprehensive assessment of the mortality from disease, injuries, and risk factors in 1990 and projected to 2020*. Boston: Harvard School of Public Health, 1996.

32. Tobias M, Christie S. Standard ICD groupers for population health: Draft 21/10/98. Wellington: Ministry of Health, 1998.

33. Clemance P. An application of correspondence analysis as a multi-dimensional scaling technique. *The New Zealand Statistician* 1985;**20**:26-34.

34. Kunst A, Groenhof F, Mackenbach J, The EU working group on socioeconomic inequalities in health. Mortality by occupational class among men 30-64 years in 11 European countries. *Soc Sci Med* 1998;**46**:1459-1476.