Public Health Monograph Series

No. 7

ISSN 1173-6844

ANONYMOUS RECORD LINKAGE OF CENSUS AND MORTALITY RECORDS: 1981, 1986, 1991, 1996 CENSUS COHORTS

NZCMS Technical Report No.3

Sarah Hill

June Atkinson

Tony Blakely

October 2002

Department of Public Health,

Wellington School of Medicine and Health Sciences

ISBN 0-473-09110-0

ISBN 0-473-09114-3 (electronic)

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

ISBN 0-473-09110-0
ISBN 0-473-09114-3 (electronic)

Acknowledgements

The Health Research Council of New Zealand is the principal funder of the NZCMS. The Ministry of Health co-funds the NZCMS.

Many, many staff of SNZ have contributed to the development of the NZCMS, with contributions ranging from quick advice to years (or more) of ongoing input into the details of record linkage. Sandra McDonald has managed the Data Laboratory and data integration aspects of the NZCMS since its inception. Valeria Kazakova, Paul Willoughby, Victoria Wilcox and Jonathan Briggs undertook the majority of the actual record linkage. We are particularly indebted to Valeria for her patience and persistence with the last two census-mortality record linkage projects and in leading the preparation of all final datasets. Keith McLeod and Victoria Wilcox contributed to the development of linkage strategies. Some (but not all) of the other SNZ staff who contributed to the NZCMS project at various stages were Robert Templeton, Tracey Gilmour, Richard Arnold, Max Wigbout, Frances Krsinich, John McGuigan, Melanie Gin, Katrina Dash, Paul Brown, Vicky Barlow, Richard Penny, Robert Didham, and Robyn Bishop. We also wish to acknowledge the managerial oversight of Sharleen Forbes, John Cornish, Dallas Welch and Brian Pink. Finally, we wish to acknowledge the essential contribution of Len Cook (former Government Statistician) who initially approved the NZCMS.

From NZHIS we wish to acknowledge the help of Tracey Vandenberg, Barbara Bridger, Liz Mooney and Jim Fraser.

Martin Tobias and Barry Borman at the Ministry of Health have provided strong support throughout.

Finally, there is a larger team of researchers and co-investigators involved with the NZCMS development than those who authored this report. Clare Salmond, Alistair Woodward, Peter Davis and Neil Pearce have all provided direct assistance at various points with the completion of the record linkage. Jackie Fawcett has provided help with some of the final stages of this report, and detailed discussions with SNZ.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Executive Summary

The New Zealand Census-Mortality Study is a population-wide cohort study, in which the cohort consists of the entire resident population and the outcome of interest is mortality. For each of the four census cohorts (1981, 1986, 1991 and 1996) the cohort population has been followed for any deaths occurring within the three years following the census. A process of anonymous probabilistic linkage has been used to link mortality records back to their corresponding census records from the previous census.

Probabilistic record linkage is a process used to link two files of records, where records in one file have a corresponding record in the other file. With the aid of sophisticated software, records from the first file are compared with records from the second file in order to find 'matching' record pairs (i.e. two records belonging to the same individual). This linkage process involves comparing items of information on one record with corresponding items on another record, looking for record pairs that contain identical pieces of information and thus are deemed to apply to the same individual.

In the NZCMS, the two files being linked were census records and mortality records from the three years following the census. The software package used to carry out this linkage was Automatch® version 4.2. The linkage process was carried out four times: once for each of the four census cohorts. Linkage of matching record pairs was made on the basis of variables included in both census and mortality files, including date of birth, country of birth, sex, ethnicity, and (most importantly) address of usual residence.

Overall, around three-quarters of all mortality records were linked with a census record. Estimates of linkage accuracy indicate a highly accurate linkage process, with an overall accuracy of over 96-98%. The proportion of mortality records linked and the estimated accuracy of the linkage process for each of the four census cohorts are as follows:

Table 1: Linkage rates and accuracy for the four census cohorts

Census cohort	*Proportion of mortality records linked*	*Estimated accuracy of linkage process*
1981	71.01%	96.9%
1986	73.78%	96.7%
1991	76.58%	98.1%
1996	78.15%	97.4%

The primary aim of the NZCMS is to determine mortality rates within different socio-economic strata of the New Zealand population, and so estimate the association between socio-economic factors and mortality. The complex process of record linkage is fundamental to achieving this aim: this technical report bears testimony to the enormous amount of time and effort that has gone into this process. Nevertheless, linkage of census and mortality data is only the first step in the Census-Mortality Study; the real work of data analysis is only now beginning. We are pleased to present the results of the

linkage process, along with the four linked datasets that are now ready for cohort analysis.

Tony Blakely

June Atkinson

Sarah Hill

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: +64-4-931 4600
Facsimile: +64-4-931 4610

Glossary

Area unit (AU)	An administrative unit referring to a geographically defined population group of around 2,000 individuals. Area units are used by Statistics New Zealand, particularly in relation to census data (thus the term Census Area Unit or CAU).
Array	Where more than one value is presented for the same variable (e.g. some mortality records contain two different dates of birth for the same individual – one from the NHI database and the other from the NMDS database).
Automatch®	A software package for carrying out probabilistic record linkage.
Bias analysis	Estimating any systematic differences between linked and unlinked mortality records (i.e. analysis of linkage bias).
Blocking variable	A variable used to break down large files into smaller subsets, to limit the number of possible comparison pairs. Comparison pairs are only formed when the blocking variable agrees exactly.
Blocks	The subsets resulting from blocking of larger files.
Clerical review	Investigator review of the records in a comparison pair, in order to decide whether or not these records are likely to apply to the same person. Clerical review usually occurs only for comparison pairs with a total weight within the cut-off range for the relevant linkage pass.
Cohort analysis	Epidemiological analysis of linked census-mortality cohort datasets to determine differences in mortality rates by social factors. (This is the primary aim of the New Zealand Census-Mortality Study.)
Comparison pair	Any possible comparison of a record from one file with a record from another file. In the NZCMS, comparison pairs consist of one census and one mortality record.
Cut-off weight	The total weight used as a threshold to decide which comparison pairs to accept as links, and which to reject. This weight is usually expressed as a discrete value, but may also be expressed as a range (where upper value = *acceptance weight*, lower value = *rejection weight*); in this case, all comparison pairs falling within the cut-off range are subjected to clerical review.

DA record	'Extra' census record from a duplicate pair – i.e., Automatch® has found two census (A) records that match the same mortality (B) record with total weight above the cut-off. One of these census records will be listed as part of a matching pair (MP), and the other as a duplicate match (DA). (The pair with the highest total weight will be listed as MP.)
DB record	'Extra' mortality record from a duplicate pair – i.e., Automatch® has found two mortality (B) records that match the same census (A) record. One of these mortality records will be listed as part of a matching pair (MP), and the other as a duplicate match (DB).
Datamail	Commercial company that specialises in geocoding.
Dataset or Database	A large collection of information files, often stored in electronic form.
Decedent	Deceased person.
Disagreement weight	See **Weight**
Domicile Code	A classification system used by NZHIS to describe geographically based administrative units. Each domicile code refers to an area containing a median population of about 2,000. The NZHIS domicile codes have a one-to-one concordance with SNZ census area units, but (unfortunately) use a different coding system (due to historical limitations in the NZHIS database).
Duplicate pair	Two records from one file, which can both, form a comparison pair with a single record from the other file, and each comparison pair has a total weight above the cut-off (i.e. both are potential links).
False negative link	A comparison pair that is not accepted as a link, but is in fact a match.
False positive link	A comparison pair that is accepted as a link, but in fact is not a match.
Frequency ratio	The ratio of the probability of variable agreement in a matching pair to the probability of variable disagreement in a non-matching pair – i.e. m / u. The frequency ratio gives a measure of the relative significance of agreement on a particular variable. It is converted to a logarithmic scale for ease of comparison (see **Weight**).
Field	The information for each variable as presented in a file. For example, the 'income' field in the census file contains the information for the variable 'income' for each record (or person). In a computerised file, fields are often represented by columns.

File	A collection of multiple records. In the NZCMS, File A refers to census records, while File B refers to mortality records.
Geocode	A code referring to a geographically based unit of administration, forming part of a classification system. Geocodes referred to in this study include area units, domicile codes and meshblocks.
Historical Mortality Data Set (HMDS)	An NZHIS dataset containing death event records for all New Zealand residents who died prior to 1988. This was replaced by the death file of the National Minimum Data Set in 1988. All NZCMS mortality records for deaths prior to 1988 are derived from the HMDS.
Linkage bias	Systematic differences by socio-demographic factors (e.g. age, deprivation) between linked and unlinked mortality records.
Links	1. A comparison pair that is accepted as being highly likely to apply to the same individual. In the NZCMS > 95% of links are probably matches. 2. A golf course (i.e. open ***field***), where ***matches*** are often played with little ***agreement*** between ***pairs*** (>95% of score cards are estimated to be correct, the remaining probably systematically biased)
MP pair	A linked (probably matching) pair of records, consisting of one census record (A) and one mortality record (B). The total weight for the pair is above the specified cut-off for the given Automatch® pass.
***m*-probability**	See **Probability**
Match	A pair of records that applies to the same individual (i.e. true links).
Match run	The sequence of passes used to link two files of records.
Matching variables	Variables common to two sets of records, for which we determine agreement or disagreement when comparing records.
Meshblock	The smallest geographic area used for coding purposes by Statistics New Zealand, with a median population size of 90-100.
National Health Index (NHI)	An NZHIS dataset, containing data for nearly every individual in New Zealand. This data is collected and updated every time a person uses public health services (e.g. outpatient visits, diagnostic investigations). The NHI dataset can be linked to NMDS events for the same individual by means of a unique identifier (the NHI number).

National Minimum Data Set (NMDS)	A dataset administered by NZHIS. Contains data for most individuals in New Zealand on both hospitalisation events and (where deceased) death events. Unlike the NHI dataset, which is updated for each new event, the NMDS contains a separate record for each hospitalisation event and thus provides several separate records for the same individual.
Non-links	A comparison pair that is *not* accepted as being highly likely to apply to the same individual.
Non-matches	Pairs of records that do not apply to the same individual (i.e. true non-links)
Partial agreement weight	The process of assigning an intermediate weight to variables that 'almost' agree (e.g. where 'year of birth' differs by only one year). This intermediate weight is less than the agreement weight but greater than the disagreement weight (thus the term 'partial agreement weight').
Pass	The process of linking two files for a given specification of blocking variable, matching variables, *m* and *u* probabilities, and cut-off weight. A series of passes carried out on the same two files is called a match run.
Positive predictive value (PPV)	The percentage of linked records that are matches (or 'true links').
Probabilistic record linkage	Record linkage of two (or more) files using the probabilities of agreement and disagreement between a range of matching variables. (This is distinct from deterministic record linkage, which links files on the basis of exact agreement between matching variables.)
Probability	
• ***m*-probability**	The probability that a matching variable agrees, given that the comparison pair in question is a match. This probability generally reflects the accuracy of the recorded data (e.g. if this is 100% accurate for both types of records, the *m*-probability will always be 1.0).
• ***u*-probability**	The probability that a matching variable agrees, given that the comparison pair in question is a non-match. This probability is generally determined by the likelihood of both records having the same value due to chance.
Random Rounded (RR)	Rounding of numerical values to the nearest multiple of three. Wherever this report refers to a particular group of census records, the total number of records will be random rounded in order to protect confidentiality.
Record	A set of variables applying to a single individual, observation or unit. In a computerised file, records are often represented by rows.

Record Linkage	The process of linking two or more files by looking for agreement or disagreement between matching variables within individual records.
Rejection weight	The total weight set as a threshold for determining which comparison pairs are *not* accepted as links (i.e. the records are deemed to apply to two different individuals).
Sensitivity	The proportion of matches detected as links, i.e. [true links] / [matches].
Skipping	Where two matching records fail to be linked because one of the records has been assigned to the incorrect block (on the basis of an erroneous blocking variable).
Specificity	Using either file in the record linkage process, the proportion of non-matching records detected as non-links, i.e. [true non-links] / [non-matches]. Note: a) the specificity varies depending upon which files it is calculated; b) the specificity can also be calculated from the perspective of comparison pairs (as opposed to records).
Total weight	The sum of the agreement / disagreement weights for each matching variable in a comparison pair of records.
True negative link	A comparison pair that is not accepted as a link, and is in fact a non-match.
True positive link	A comparison pair that is accepted as a link, and is in fact a match.
***u*-probability**	See **Probability**
Value-specific weightings	Agreement and disagreement weights that are specific to the actual value of a given variable. Value-specific weightings are used where some values are far less common than others, so the relative significance of an agreement for that value is much greater. For example, the agreement on New Zealand as country of birth adds much less weight than an agreement on Africa.
Weighting	The process of assigning a value to all possible comparisons of matching variables.
Weight	
•**Agreement weight**	The value assigned for agreement on a given matching variable. This value is a positive number, calculated from the *m* and *u* probabilities for that variable according to the following formula: $[\ln (m / u) / \ln(2)]$.

•**Disagreement weight**	The value assigned for disagreement on a given matching variable. This value is a negative number, calculated according to the following formula: [ln ((1-*m*) / (1-*u*)) / ln(2)].

Abbreviations

AU	area unit (median population about 2,000)
CAU	census area unit - i.e. an area unit derived from census data
dd	day of birth
E[FP]	Expected number of false positive links
HMDS	Historical Mortality Data Set
mm	month of birth
NHI	National Health Index
NMDS	National Minimum Data Set
NZCMS	New Zealand Census-Mortality Study
NZHIS	New Zealand Health Information Services
NZSCO-68	New Zealand Standard Classification of Occupations, 1968
NZSCO-90	New Zealand Standard Classification of Occupations, 1990
NZSEI	New Zealand Socio-Economic Index (an occupational class index)
PPV	Positive predictive value
SNZ	Statistics New Zealand
yyyy	year of birth

Chapter 1 Introduction

The primary aim of the New Zealand Census-Mortality Study is to determine mortality rates within different socio-economic strata of the New Zealand population, and so estimate the association between socio-economic factors and mortality.

This is being undertaken through a series of cohort studies, where the cohort consists of the entire New Zealand population and the follow-up period is the three years following each census. The exposures of interest are socio-economic factors, and the outcome of interest is death in the three years following census night. Thus, in calculating mortality rates, the numerator (number of deaths) is derived from mortality data, while the denominator (population number) is derived from census data.

If socio-economic factors were included on mortality records, it would be possible to calculate stratified mortality rates using unlinked census and mortality data. Unfortunately this is not the case: detailed socio-economic data (including education, labour force status, car access, housing tenure and household income) is included in census records, but not in mortality records. Therefore, the only way to calculate stratum-specific mortality rates is to link each mortality record back to its corresponding census record. This allows us to unite each decedent's mortality record with the socio-economic data recorded on the corresponding census form, and so assign each death event to the appropriate socio-economic stratum.

The possibility of linking New Zealand mortality and census data was initiated by Tony Blakely in 1996. The first population cohort to be linked with mortality records was that from the 1991 census. Further detail on the 1991 census-mortality linkage project is provided in the following documents:

- detailed documentation of the 1991 census-mortality linkage process and outputs (Blakely et al. 1999; Blakely et al. 2000)
- development of methods to determine the accuracy (positive predictive value) of the record linkage(Blakely and Salmond in press; Blakely et al. 1999)
- linkage bias analyses (Blakely et al. 1999)
- analyses to unlock the numerator-denominator bias for ethnicity between census and mortality data (Blakely and Atkinson 2001; Blakely et al. 2002a; Blakely et al. 2002b)
- cohort analyses of adult mortality (Blakely 2002)(Blakely 2001; Blakely et al. 2002c).

Following the success of the 1991 cohort study, linkage was carried out for three other census cohorts: 1981, 1986 and 1996. The process of probabilistic record linkage was developed during each of these projects, with constant learning and refining of linkage techniques. During this process it became apparent that the NZCMS offered valuable information on many areas besides socio-economic determinants of mortality.

One very salient finding is that systematic misclassification of ethnicity has occurred on many mortality records, with many Maori and Pacific decedents mistakenly classified as

non-Maori non-Pacific. This misclassification means that mortality rates for Maori and Pacific Islanders have historically been underestimated: numbers of Maori and Pacific decedents (the numerator) have been underestimated compared with numbers of Maori and Pacific residents (the denominator), which are obtained from self-identified ethnicity on the census form. This phenomenon is referred to as 'numerator-denominator bias'. The discovery and quantification of bias in historical estimates of Maori and Pacific mortality rates has significant implications for future policy decisions.

> The objective of this report is to describe the linkage methods, data requirements, linkage process, linkage outputs, and main analytical files for the four census-mortality projects in the NZCMS.

By linkage *methods* we mean:
- the anonymous and probabilistic record linkage methodology;
- methods developed in the NZCMS to determine the accuracy of the linkage.

By *data requirements* we mean the preparation of mortality and census files for linkage, including descriptions of the necessary variables and geocodes.

By linkage *process* we mean the steps we undertook to link each of the four censuses to mortality data, including the estimates of linkage accuracy.

By linkage *outputs* we mean the three files directly arising from the record linkage (linked census-mortality records, residual mortality file and residual census files), including the numbers of records in each file.

By *main analytical files* we mean the bias, unlock, and cohort files for each of the four census-mortality projects (i.e. 12 files in total).
- The bias file for each census-mortality project consists of all mortality records for that project, with an indicator variable for whether the mortality record was linked to a census record. These bias files allow us to determine the differences in demographic and other characteristics between those mortality records linked and unlinked (i.e. *linkage bias*).
- The unlock files consist of the subset of highly probable linked census-mortality records. We have used these files to determine the discrepancy between ethnicity recorded on census versus mortality data (i.e. unlocking the so-called numerator-denominator bias that plagues all routine calculations of mortality rates by ethnicity in New Zealand.)
- The cohort files consist of the full census files with information on mortality for those census records linked to a mortality record. (The cohort files have been weighted to adjust for linkage bias.) The cohort files are the major analytical files in the NZCMS that will be used for the majority of research outputs in the future.

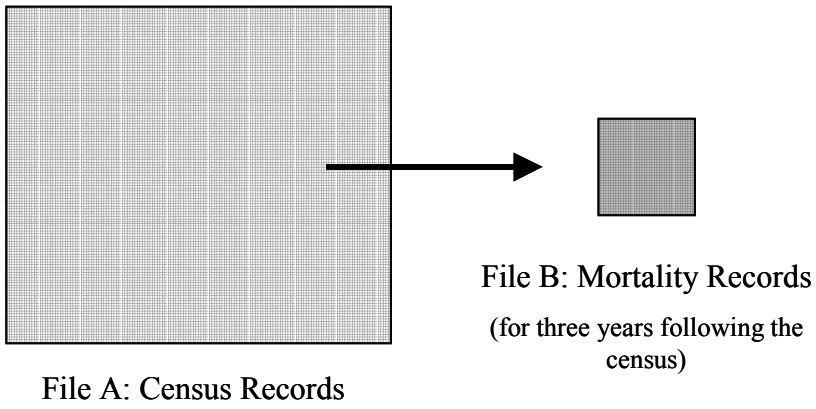
This report does not describe the unlock analyses, analyses of linkage bias or cohort analyses – these activities are described in accompanying technical reports (Ajwani et al. 2002; Fawcett et al. 2002). Rather, this Report stops at the point of describing the content of the three files necessary for this work.

The NZCMS offers a wealth of data on the health of the New Zealand population. We hope that this study will become a useful source of information for many researchers, government departments and other organisations involved in promoting health and reducing inequalities on a population-wide basis.

Chapter 2 Methods

We start with two sets of records: File A and File B. These two files contain records for the same population. In our case, File B consists of mortality records from a three-year period, and File A consists of records from the census immediately preceding that period. Thus (in theory) each record in File B has a corresponding record (belonging to the same person) in File A.

Figure 1: File A (census records) and File B (mortality records)



We want to match up those records from File A and File B that belong to the same person. Each record contains various pieces of information about an individual, but there is no name or unique identification code that will allow us to directly match records from the two files. Instead, we will have to match records on the basis of the information they contain.

Humans searching two files for the same individual intuitively do two things. First, they look for agreement or disagreement on variables included in both files (***matching variables***). Second, they assign varying importance to different variables. For example, a match on an NHI number (or some other unique identifier) just about guarantees the records in the two separate files are for the same person. But a match on sex adds only a small amount of discriminatory information.

2.1. Probabilistic Record Linkage

Probabilistic record linkage formalises these intuitive processes by:

- Looking for agreement or disagreement between the matching variables contained in two records ***(record linkage)***.
- Assigning relative significance to each agreement/disagreement on the basis of probabilities ***(weighting)***.

The combination of these two processes constitutes probabilistic record linkage. The NZCMS is based on probabilistic record linkage using Automatch®, a software package developed by Jaro.(Jaro 1995)

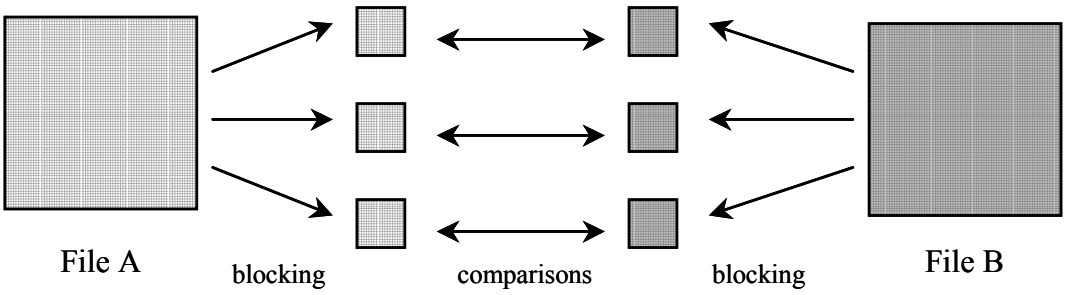
2.1.1 Blocking

Comparing records from two large files requires an enormous number of comparisons. For example, if two files contain 1,000 records each, there are 1,000 x 1,000 = 1,000,000 possible ***comparison pairs***. Such a large number of comparisons is computationally inefficient, and is likely to produce a number of false links (i.e. incorrect links that occur purely by chance). The first step in record linkage is therefore to break down each large file into smaller subsets, which are then compared.

Both files (A and B) are broken down into smaller subsets using one of the variables contained in both files. For example, records might be grouped according to a person's year of birth. The variable used is called the ***blocking variable***, and the subsets that result from it are known as ***blocks***.

Records within a given block on File A are then compared to all records in the same block on File B, looking for agreement/disagreement on matching variables (such as gender, day and month of birth). Since each block contains a relatively small number of records (compared with the whole file), a much smaller number of pair combinations are possible, and record linkage is therefore more efficient.

Figure 2: Blocking

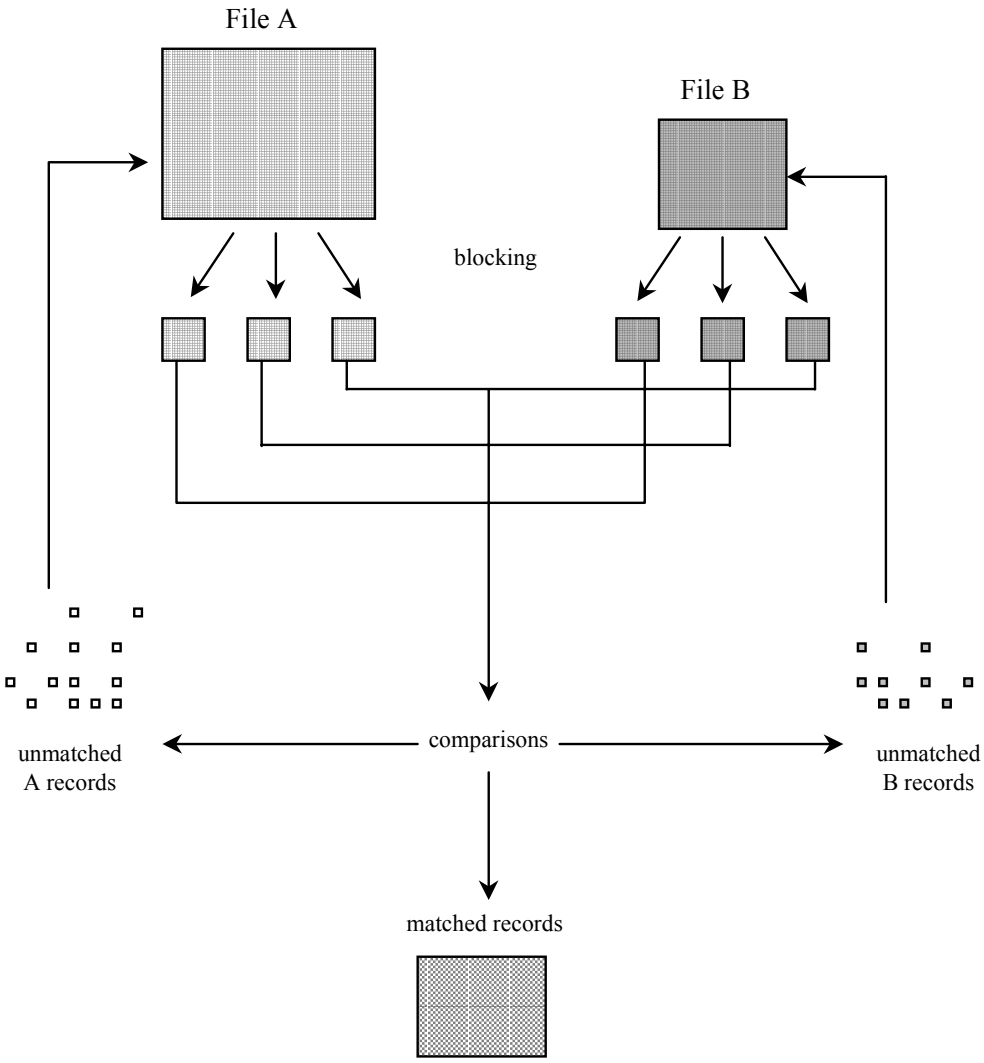


A problem with blocking is that it relies on both records of a true pair (i.e. a ***match)*** containing correct data for the blocking variable being used. If the blocking variable is incorrectly entered on one record, this record will end up in the wrong blocking subset and cannot be matched with its pair. This problem is known as ***skipping***.

In order to avoid non-matching due to skipping, blocking of files should ideally be carried out several times using a different blocking variable each time. The first blocking variable is chosen, and records are subdivided accordingly. The matching process is then undertaken, and some (hopefully most) of the records are matched on this basis. Those records that have not been matched are then pooled together, and a second blocking variable is chosen. Since this second blocking variable is different to the first blocking variable, records assigned to the wrong subset (due to an erroneous variable) during the first block now have the opportunity to be assigned to the correct block. Thus records that failed to be matched in the first matching process may be matched in the second, and so on.

This process is repeated several times, using a different blocking variable for each matching process. Each time the number of unmatched records remaining becomes smaller, until eventually the marginal return is negligible.

Figure 3: The match run process



A single matching process (using one blocking variable) is called a *pass*. The sequence of passes used to match two sets of records is known as a *match run*.

2.1.2 Weights

Matching has greater significance for some variables than for others. For example, agreement on sex has relatively little significance, whereas agreement on date of birth has more significance. In probabilistic linkage, we account for this by assigning weights to the different variables being matched.

The *m probability* is the probability of agreement for a given variable between two records for the same individual (i.e. a matching pair). The *u probability* is the probability of agreement between records for two different individuals (i.e. a non-matching pair).

The *m* probability depends upon the accuracy of the recorded data. Where two records are for the same individual they should (in theory) match for every variable, in which case the m probability would be 1.0 (100%) for every variable. In practice, however, there are occasional non-agreements between records for the same person due to errors in recording of data or differences in the way the data was obtained. This results in an *m* probability less than 1.0. For example, if day of birth is incorrectly recorded in 5% of records, the m probability for day of birth will be 0.95.

The *u* probability depends mainly on the likelihood of a variable matching due to chance. For example, a person's day of birth will be one of 31 possibilities. Thus the likelihood of a match for day of birth between two different individuals is 1/31 or 0.032.

The two probabilities (*m* and *u*) are used to give the *frequency ratio* for the variable in question. This is a measure of the relative significance of agreement for this particular variable. The agreement frequency ratio is equal to the probability of a match between two true links divided by the probability of a match between two non-links (i.e. *m* / *u*).

Frequency ratios can also be calculated for *disagreement* between two variables. This gives a measure of the relative significance of a non-match for this variable. The disagreement frequency ratio is equal to the probability of a non-match between two true links, divided by the probability of a non-match between two non-links (i.e. [1-m]/[1-u]).

From the above we can see that agreement frequency ratios are expressed as numbers of increasing size from 1 to + ∞, whereas disagreement frequency ratios are expressed as fractions between 0 and 1. For ease of use it is conventional to convert these ratios to a linear scale using the natural logarithm to base two. (This log transformation also means that weights can be summed for each matching variable comparison, rather than multiplying the frequency ratios.) An example of this (including the formula) is given in Table 1. Using logarithms to base two not only converts the frequency ratios to a linear scale, but also means that agreement ratios are expressed as positive numbers, while disagreement ratios are expressed as negative numbers. Thus we have created a scale for measuring the *relative probability of a true match*, where positive numbers represent increasing probability and negative numbers represent decreasing probability. This relative probability is called the *weight*. (Note, however, that this weight does not correspond to the *actual* probability of a comparison pair being a match. Rather, one can compare weights for different comparison pairs to determine their *relative* likelihood of being a match.)

Table 2: Example of agreement and disagreement frequency ratios and weights for comparison by matching variable 'day of birth'

Comparison Outcome	Proportion Links	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.

Frequency ratios and weights can be calculated (as above) for each of the variables contained in a set of records. We take a record from each of two files and compare the two looking for (dis)agreement on each of the matching variables (e.g. day, month and year of birth, country of birth, ethnicity). Using our calculated frequency ratios we can then assign a weight for each variable, depending on whether it matched ($\rightarrow$ agreement weight) or did not match ($\rightarrow$ disagreement weight). We then add together the individual weights to give a *total weight* for this comparison pair. This total weight indicates the relative probability that the two records belong to the same individual.

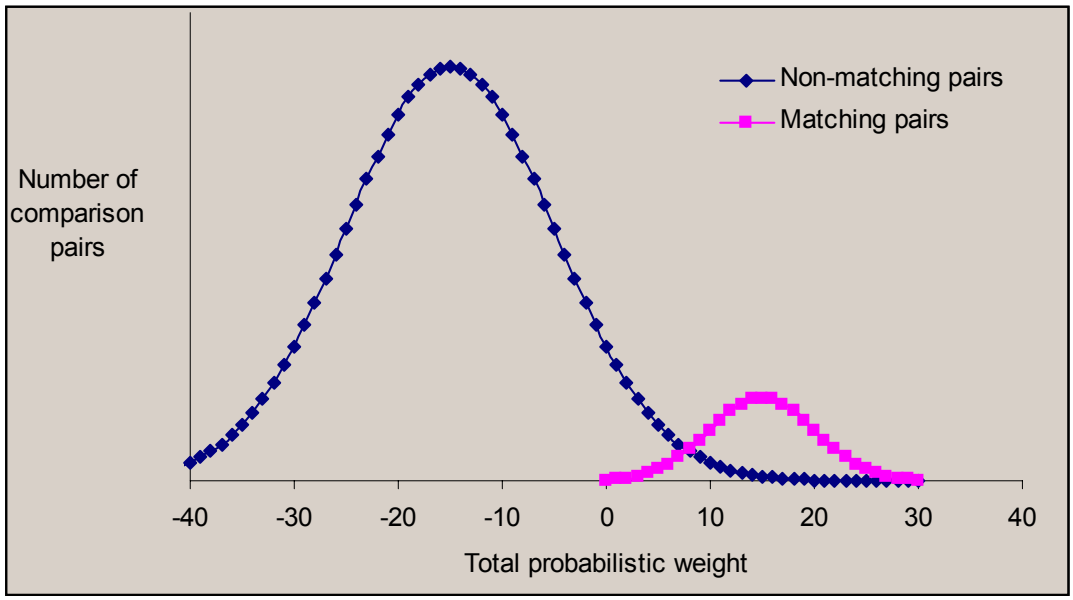
Automatch® undertakes the mechanical aspects of the weighting and linkage processes described above. However the user must first specify the *m* probability and other particulars of the matching process to be used for each variable. Details of these specifications are given in the Technical Report for the 1991 Census Cohort.(Blakely et al. 1999)

2.1.3 Determining Cut-Off Weights

When two sets of records are compared with each other, Automatch® calculates a total weight for every possible comparison pair. The majority of comparison pairs will consist of records from two different individuals (*non-matching pairs*). Since these records will disagree on most variables, their total weight will be highly negative. Conversely, a smaller number of pairs will contain records from the same individual (*matching pairs*). Since these records agree on most variables they have a highly positive weight. Thus the distribution of the total weights for two sets of records is generally bimodal (see Figure 4).

A few pairs will have a total weight in the intermediate range. This indicates a match for some variables, but not for others. A number of these pairs will represent true links (with some variable disagreement due to error or differences in data recording), while others will be false links (with occasional variable agreement due to chance).

Figure 4: Distribution of matching and non-matching pairs by total probabilistic weight



The total weight gives us an indication of how sure we are that a comparison pair is a match (or true link). We can use this total weight to decide which pairs to accept as links, and which to reject. The cut-off threshold we choose represents a trade-off between the *number* and the *accuracy* of the links obtained. A higher threshold means we accept only pairs we are very sure to be true links, which gives us only a small number of links. Conversely, a low threshold means we accept many more pairs as true links, but a number of these are likely to be false links (see Section 2.3 below).

We may be able to improve both the accuracy and the yield of our linkage by undertaking *clerical review* of those comparison pairs with an intermediate total weight. This improvement arises from our understanding of coding practices that we have been unable to quantify probabilistically using Automatch®. For example, we know about systematic differences between datasets in how ethnicity is recorded which allow us to be more discriminatory in clerical review than is possible when relying on a computerised decision.

To include clerical review in our linkage process we set two threshold weights: a relatively high *acceptance weight*, and a relatively low *rejection weight*. All those matches whose weight lies in between these two cut-offs are then subjected to clerical review in order to decide whether the match should be accepted or rejected. (NB: 'Clerical review' refers to review of the relevant variables on the electronic records that were entered into the Automatch® database. It does not involve review of any physical records.)

2.1.4 Improving Discriminatory Power

The methodology described thus far gives an overview of the basic principles involved in probabilistic record linkage. The use of more sophisticated techniques allows us to

improve the discriminatory power of our record linkage. These techniques include use of value-specific weightings, partial agreement weights and arrays of variables.

2.1.4.1 Value-Specific Weightings

We have seen how *m* and *u* probabilities are used to generate weightings for particular variables. The method described above gives a standard agreement or disagreement weight for each variable, regardless of the actual value of that variable.

Some values, however, are much less common than others, and a match between two records is therefore more significant. For example, a birthplace of 'Australia' is relatively common amongst New Zealand residents; thus a match for this variable (due to chance) is far more likely than for the birthplace 'Sweden'.

This differential significance can be utilised by assigning different weightings to specific *values* for the same variable. The use of value-specific weightings increases the discriminatory power of the record linkage process. In the NZCMS, specific *m* and *u* probabilities were particularly relevant for the variables of ethnicity and country of birth.

2.1.4.2 Partial Agreement Weights

Records are compared with each other in terms of corresponding variables. Where these variables agree, an agreement weight (positive) is assigned, and where they disagree, a disagreement weight (negative) is assigned.

Sometimes numerical variables are very close to agreeing, and differ by only a single digit. For example, it is common for the year of birth to be reported and entered incorrectly by only one or two years. An absolute difference in year of birth of only one or two years is not as bad as a difference of, say, 30 or 50 years. It may therefore be appropriate to assign a partial agreement weight for those variables that are very close to agreement.

Automatch® can be used to assign partial agreement weights using the PRORATED function. This allows variation in numerical matching by a specified amount. For example, the variable 'year of birth' can be matched with a specified tolerance of two years. An exact match for year of birth will still be assigned the full agreement weight. A mismatch by one year will be assigned an intermediate weight, one third of the way between a full agreement weight and a full disagreement weight. A mismatch of two years will be assigned a weight two thirds of the way towards a full disagreement weight.

2.1.4.3 Arrays

An array is where more than one value may be used for the same variable. For example, some mortality records contained two different dates of birth for the same individual. This is because information was often obtained from more than one source (such as both NZHIS and NMDS databases), and occasionally these two sources would differ in the information they contained for the same person. Obviously at least one of these values must be incorrect, but we have no way of determining which is the true date of birth for that person.

Since we don't know which variable is correct, it is useful to be able to use both variables when carrying out record linkage. This maximises our chance of finding the corresponding census record for that person. Automatch® allows variables to be specified as arrays. This results in:

- a *full agreement weight* if either or both values on the mortality record *agree* with the census variable[1]
- a full disagreement weight if both values on the mortality record disagree with the census variable

In practice, Automatch® version 4.2 was unable to simultaneously compute value-specific weightings, partial agreement weights and arrays for the same variable. The actual specifications we used for each census cohort are described in Chapter 4 of this Report.

2.2. NZCMS Record Linkage Strategy

There are no hard and fast rules about which variables to use for blocking and which to use for matching, or the order in which passes should be undertaken. The aim of the linkage process is to link as many records as possible as accurately as possible. The best blocking variables and sequence to use for this are generally determined *a priori* with modification by trial and error.

In general, it is best to begin by dividing the files into the smallest blocks possible, as this limits the number of comparisons to be undertaken for each block. This means selecting the variable that has the greatest number of possible values. For example, blocking a file according to month of birth would produce twelve fairly large blocks, whereas blocking according to day of birth would produce 31 much smaller blocks.

In the NZCMS, the variable that gave the smallest blocks was *meshblock of residence*. ('Meshblock' refers to a small geographical area containing approximately 100 people.) Meshblock was therefore used as the initial blocking variable for all record linkage. A process of trial and error determined subsequent blocking variables and the matching variables used in each pass. Much time was spent determining which combination of variables and which sequence of passes gave the best results for each census cohort. This report presents only the final match run strategy that was used for each census cohort (see Chapter 4: Record linkage process and outputs).

2.2.1 Automatch® Terminology

The NZCMS uses two files for each census cohort. File A consists of all census records for people aged 0-74 years on census night. File B contains mortality records for all people aged 0-74 on census night who died in the three years following the census.

[1] Ideally, we would have preferred less than a full agreement weight if only one of the members of the array agreed. However, this was not possible with Automatch® using the specifications we required.

Theoretically, every record in File B should have a corresponding record in File A. (In practice there are some mortality records for which there is no corresponding census record – e.g. where the decedent was not in New Zealand on census night or did not fill out a census form.)

When Automatch® finds two records with a total weight above the chosen threshold, these are accepted as a link and are designated 'MP' (matched pair).

Sometimes Automatch® will find two census records that match a single mortality record, both with probability weights over the threshold. In this case, the pair with the highest weight will be designated 'MP', and the remaining census record will be designated 'DA' (duplicate A). (Where both pairs have the same weight, Automatch® arbitrarily assigns one as 'MP' and the other as 'DA'.)

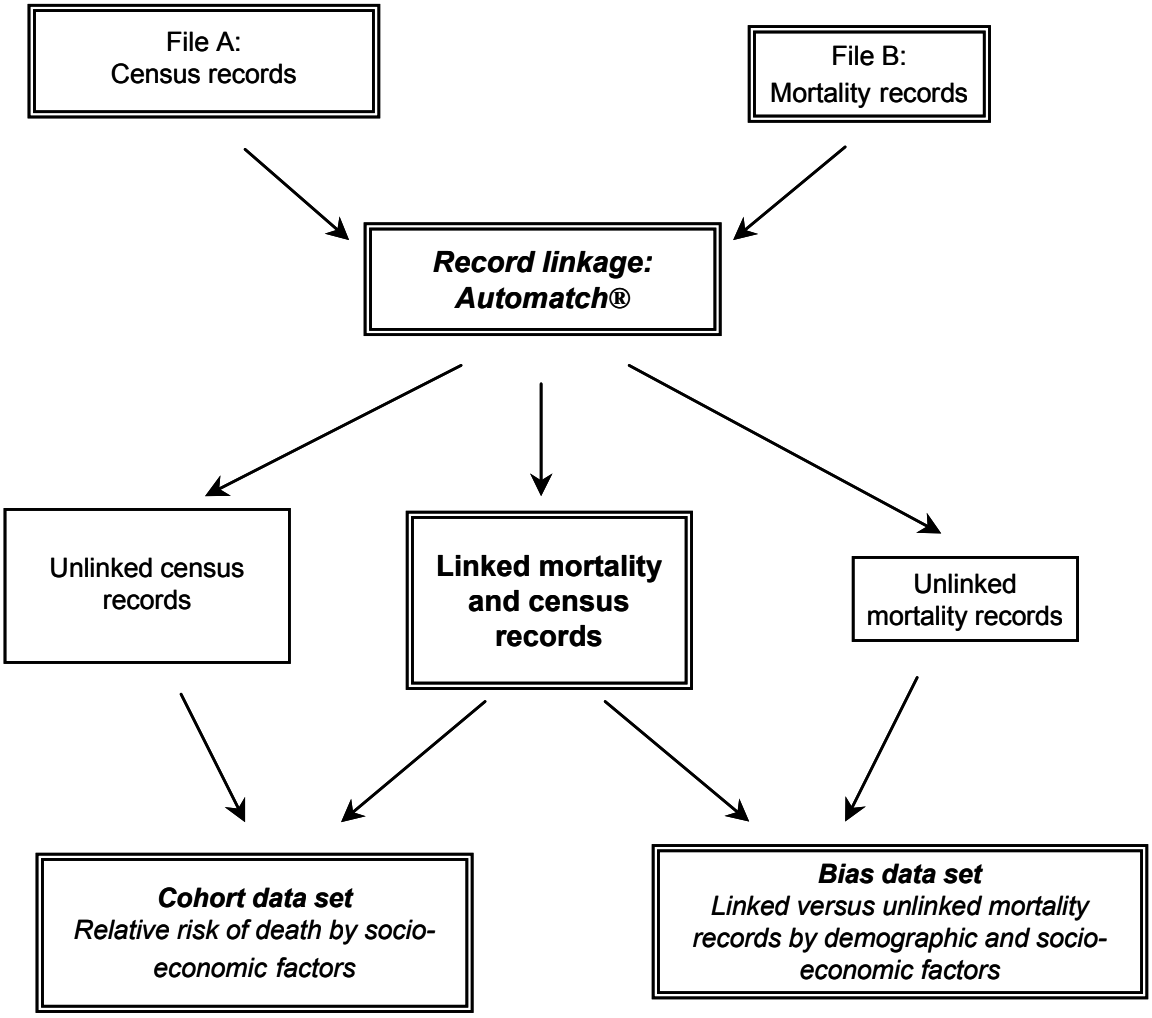
Less commonly, two mortality records will be matched with the same census record. Again, the pair with the highest weight will be designated 'MP', and the remaining mortality record categorised 'DB' (duplicate B).

DA and DB records are kept 'associated' with their corresponding MP pair for later clerical review. If clerical review confirms one combination as the most likely match, this pair is accepted as the true link and the remaining unpaired record is returned to the residual unmatched file. If neither combination is distinguished as the most likely, both possible pairs are rejected and all three records treated as unmatched.

In practice, around three-quarters of all mortality records were successfully matched with a census record and accepted as links. The remaining unmatched mortality records were stored separately, and used later in estimating the degree of bias involved in the linkage process (*bias analysis*).

Only a very small proportion of individuals who completed census forms would be expected to die in the following three years. Thus the vast majority (~99%) of census records (File A) remained unlinked at the end of the linkage process.

Figure 5: The record linkage process



2.3. Determining the Accuracy of the Record Linkage

The very nature of probabilistic record linkage means we cannot ever be entirely certain that a linked pair of records do actually belong to the same person. Hopefully the (vast) majority of links are matches (**true positives**), and only a few matches are missed (**false negatives**). However we do need some way of estimating the accuracy of our record linkage in order to validate our study results.

A two-by-two table of link/non-link status by match/non-match status is shown below. If the outcome was death, then matches would be those who died during follow-up and non-matches would be those alive at the end of follow-up.

	Matches	Non-matches
Linked	**a** (true positives)	**b** (false positives)
Unlinked	**c** (false negatives)	**d** (true negatives)

Note that the record linkage is acting like a screening or diagnostic 'test' for the actual outcome/match status, by categorising comparison pairs of records as either linked (test positive) or unlinked (test negative). In probabilistic record linkage this categorisation is achieved by setting a cut-off score above which comparison pairs are considered linked, and below which they are considered unlinked. (The higher the cut-off, the more probable the link is to be a match). Accordingly, we can quantify the performance of the record linkage in *classifying the outcome* with the familiar terms:

Sensitivity	= a / (a+c)
Specificity	= d / (b+d)
Positive predictive value	= a / (a+b)
Negative predictive value	= d / (c+d)

These parameters will vary depending on the cut-off: decreasing the cut-off will increase the sensitivity, but also increase the number of false positives; increasing the cut-off will decrease the sensitivity, but also decrease the number of false positives.

Since we have no 'gold standard' for establishing which matches are true links, we must use indirect measures to estimate the sensitivity and specificity of our linkage process.

The *sensitivity* of our record linkage may be estimated from the proportion of mortality records that were successfully linked in the final match run – i.e.:

$$\text{Sensitivity} = \frac{\text{Number of mortality records that were successfully linked } \mathbf{(a)}}{\text{Number of mortality records for which a link is possible } \mathbf{(a + c)}} \approx \frac{\text{Accepted links}}{\text{All true links}}$$

This estimation is based on two assumptions:

1. the number of false positive links is negligible compared with the total number of accepted links (numerator bias)
2. the number of mortality records for which there are no corresponding census records is negligible (denominator bias)

There is no corresponding simple approximation for estimating the *specificity* of record linkage. The specificity varies depending on whether it is calculated from the perspective of the mortality or census records. Unlike the sensitivity, the specificity cannot be directly estimated from the numbers of records linked.

A more useful measure than the specificity is the positive predictive value (PPV) of the record linkage. Newcombe (Newcombe 1988) describes a method based on the probabilistic weights (the 'absolute weight method') that could be used to estimate the PPV. However, this method is prone to bias. Two other methods ('chance method' and 'duplicate method') for estimating the PPV were developed specifically for this project. Details of these are provided in the first Technical Report (Blakely et al. 1999) (pages 43-61) and elsewhere.(Blakely and Salmond in press) Both these methods are applicable only to record linkage projects where there is only one true link for each record (so-called 'best linkage').

The chance method estimates the number of false positive links among exactly matching pairs. For example, on average there are about ([sex (2)] × [dd (30)] × [mm (12)] × [yyyy (60] × [ethnicity (1.2)] × [country of birth (1.2)] = 10,368 ≈ 10,000 possible combinations of exact agreements.[2] Thus, each mortality record has a 0.00001 probability of agreeing exactly with any *one given* census record. For a meshblock pass where each mortality record is compared to 100 census records, each mortality record has a 0.001 or 0.1% chance of forming a false positive link.

The duplicate method utilises the varying probability of a false positive link by total weight score in the record linkage and the occurrence of duplicate links. Its main advantage over the exact method is its applicability to non-exact agreements. Using the probabilities of a single mortality record being matched with zero, one, or two census records, it estimates the number of false positive links using binomial combinatorial probabilities. (This method is described in detail elsewhere.(Blakely and Salmond in press))

Each of the above methods has its own advantages and limitations. In practice, both methods were used to calculate the expected number of false positives (E[FP]) and the positive predictive value (PPV) for each record linkage pass. Results were similar, and indicated an overall PPV of around 97.5 – 98% for all record linkage.

[2] Where 60 for year of birth and 1.2 for ethnicity and country of birth are 'weighted' numbers given the uneven distribuition by values for these variables.

Chapter 3 Data used in the record linkage

The first step in record linkage is to obtain the two files to be linked, and define the matching variables. The two files used in the NZCMS consisted of census records (File A) and mortality records (File B). For each of the four census-mortality linkage projects the two files were put together slightly differently and required different kinds of preparation in order to be suitable for Automatch® linkage. However, the general schema was the same throughout the study. This chapter describes the general principles of this process first, followed by important specifics for each year (1981, 1986, 1991, and 1996). A description of all variables used from census and mortality data in cohort, unlock and bias analyses is presented later in 'Chapter 5: Cohort, bias and unlock'. *The present chapter focuses only on the variables necessary for the record linkage.*

3.1. Census Data

All census data is stored by Statistics New Zealand, and is kept under conditions of strict privacy. Since this data is not permitted to leave SNZ, the census files for this study were all prepared by SNZ staff. The actual record linkage process was also carried out at SNZ, so the census records stayed within the Department throughout the entire study.

Complete records from each census are kept on master-files within SNZ. The census data needed for the NZCMS were extracted from each master-file to form a smaller file for each of the four censuses (1981, 1986, 1991 and 1996).

3.1.1 Variables used in Census Records

Variables included in this file were those that would be used for integrating with mortality files (i.e. record linkage), and also those to be used later in analysing mortality rates for different socio-economic groups (i.e. cohort analysis). Some variables were also used in the 'unlock' analyses. Census variables used in the record linkage are presented in Table 3. (The full list of all census variables available for the cohort analyses is presented subsequently in Table 17, page 83).

Table 3: Census variables included for use in record linkage

Variable	Purpose	Comments
Date of Birth	• 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 mortality variables during the record linkage.
Sex	• matching variable	-
Ethnic Group	• matching variable	Each individual could nominate multiple ethnic groups on their census form. In order to allow matching with mortality records (which in most cases allowed for only one ethnic group per person), census ethnicity was usually reduced to a single ethnic group using a ranking of: 1. Maori, 2. Pacific Island and 3. non-Maori non-Pacific (i.e. all other ethnic groups). However, there was variation in how ethnicity was treated between the four census-mortality projects.
Country of Birth	• matching variable	Country of birth was categorised as one of nine values.
Meshblock	• blocking variable	The usual residence meshblock was the most important blocking variable as it has the most number of values.
Area Unit	• blocking variable	About 10% of mortality records had no meshblock, only a usual residence area unit. Area unit was therefore included in the census file to allow linking of these individuals. Also, if either the census or the mortality record for a particular decedent had a miscoded meshblock, but miscoded to another meshblock in the same area unit, then blocking by the area unit may result in a correct link.

3.1.2 Notes on Specific Variables

3.1.2.1 Ethnic Group

The New Zealand census form has elicited ethnicity in different ways over time. This has influenced the way individuals report their ethnicity in each census.

In order to use ethnicity as a matching variable, we want to define it in the same way for both census and mortality records. It is therefore necessary to have some appreciation of how ethnicity was defined in each census used in the NZCMS.

The table below outlines the way ethnicity was defined in each census, and also in the corresponding mortality records. It can be seen that the census allows each person to identify with multiple ethnic groups. This contrasts with mortality records, which (prior to 1995) recorded only a single ethnic group for each decedent.

Year	Census Data	Mortality Data
1981	~ 'biological' ethnic origin ~ multiple groups allowed	*Prior to September 1995:* ~ biological race ~ only three categories:
1986	~ self-identified ethnic origin ~ multiple groups allowed	Maori, Pacific, non-Maori non-Pacific
1991	~ self-identified ethnic origin ~ multiple groups allowed	
		After September 1995:
1996	~ self-identified ethnicity ~ more encouragement of multiple groups	~ multiple ethnic groups (identical questions to 1996 census)

Census ethnicity: 1981, 1986 & 1991

Because of the single ethnic grouping on pre-1995 mortality records, it was necessary to reduce all census records for this period to just one of three ethnic groups (corresponding with those recorded on mortality forms). These were: Maori, Pacific and non-Maori non-Pacific. It was therefore necessary to rank census ethnic groups to determine which would have dominance. The hierarchy used was: 1. Maori, 2. Pacific, 3. non-Maori non-Pacific. This immediately introduced an element of bias, as individuals with more than one ethnic group on their census record were automatically assigned to 'Maori' or 'Pacific Islander', even if their 'first choice' ethnic group would have been European. Unfortunately this was unavoidable, due to the different ways of recording ethnicity on census and mortality forms.

Census ethnicity: 1996

The ethnicity question on the 1996 census used slightly different ordering and wording to previous censuses, which had the effect of encouraging respondents to identify themselves as belonging to more than one ethnic group. (Individuals were able to identify with more than one group in previous censuses, but tended not to do so.) At the same time, a change had occurred in the collection of ethnicity data on the death registration form in September 1995, meaning that mortality records for those individuals in the 1996 – 1999 cohort included up to three ethnic groups.

For these reasons, ethnicity data was characterised differently for the 1996-99 mortality records (see description later). However, the census ethnicity variable was still characterised using the hierarchical model described above, with the addition of one other category - 'Asian' - between 'Pacific' and 'non-Maori non-Pacific'.

3.1.2.2 Meshblocks and Area Units

The smallest administrative unit used by Statistics New Zealand is the 'meshblock', which has a median population of about 100 people. On the census file, each person is coded as belonging to (residing in) a particular meshblock. The next level of aggregation

is called an 'area unit', with a median population of about 2,000. Area units used in the census are sometimes referred to as Census Area Units (CAU). Since meshblocks are smaller than area units they provide more discrimination in terms of record linkage, and are therefore preferred over area units.

A meshblock of residence is recorded for all census records, but was sometimes unable to be assigned to mortality records. Mortality records do, however, include at least one domicile code (a one-to-one concordance with area unit) which could be used as a 'second choice' measurement of location when meshblock was missing. (See later section for details on how meshblock and area unit codes were assigned to mortality records). In order to allow matching to take place in the absence of meshblock codes, all census records were also assigned their usual residence area unit code.

Census meshblock and area unit codes tend to evolve over time, as populations expand or diminish and the area boundaries are changed accordingly. All historical census records are automatically updated (by SNZ) so that area codes use the most up-to-date coding system. At the time of the NZCMS, the 1981, 1986 and 1996 census databases were using 1996 geocodes (i.e. they were all 1996-base). The 1991 census database uses the 1991 geocodes.

3.2. Mortality Data

For each census cohort, a file of mortality records was obtained for decedents who were aged 0 – 74 years on census night and who died in the subsequent three years. These records were provided by the New Zealand Health Information Service (NZHIS). The structure of the mortality file differed slightly for each census cohort, due to changes in the way data has been collected over time.

3.2.1 Sources of Data for Mortality File

Mortality data is collected by New Zealand Health Information Services (NZHIS). Since 1988, this data has been recorded in two databases: the National Hospital Index (NHI) dataset, and the National Minimum Data Set (NMDS).

3.2.1.1 National Hospital Index Data Set

The National Hospital Index (NHI) file holds demographic information for each person who comes in contact with health services based in a public hospital. Each time an individual uses these services, a hospital clerk collects personal details, which are entered into the NHI database. This is updated with each new contact; historical data is not retained.

Information contained in the NHI database includes name, address, date of birth, ethnicity and NHI identification number.

3.2.1.2 National Minimum Data Set (NMDS)

The NMDS contains records for a number of health events relating to any one person, including (in the cases of deceased persons) the death event. Each record contains personal information as well as clinical details relating to the health event (e.g. ICD code). The actual address for the person is not entered; instead, a computer-generated domicile code is entered for their residential address at the time of the health event.

Regarding the NMDS death event, it is important to note that the demographic details (DOB, sex, ethnicity and country of birth) are usually entered independently of the NHI file. The information is elicited by an undertaker and entered on the death registration form (BDM28). This form is sent to SNZ, who enter the information it contains and pass this on to NZHIS. The only situation in which NHI and NMDS data are derived from the same source is when a decedent has no previous hospitalisation event, in which case the NMDS demographic data is used to construct the NHI file.

NMDS records include the NHI number so that several hospitalisations and other NMDS events can be identified as arising from the same person.

3.2.1.3 Historical Mortality Data Set

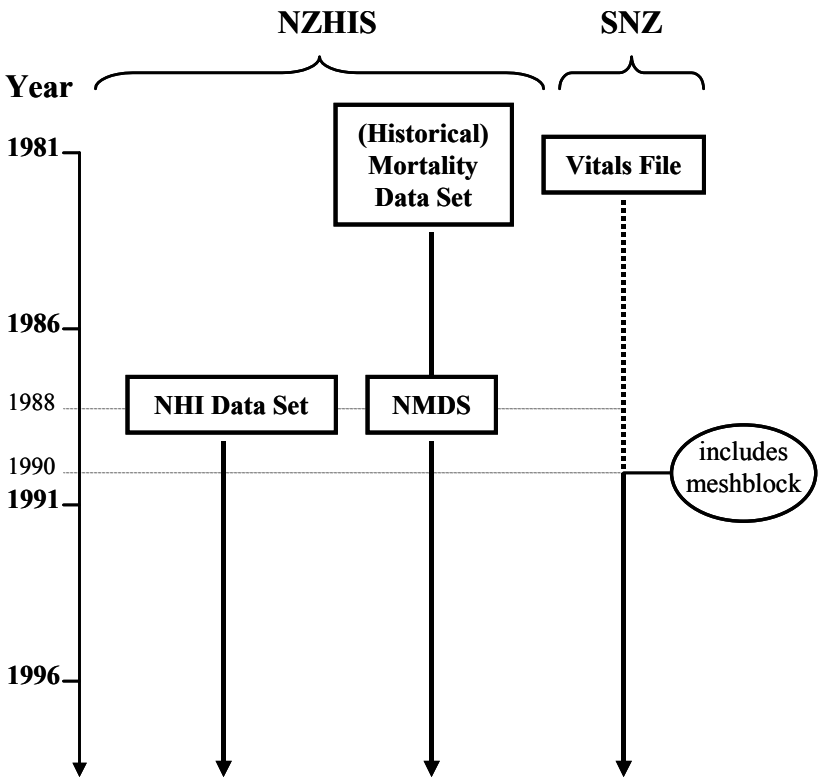
Prior to 1988, NZHIS kept death event data in a Mortality Data Set. After the introduction of the NHI system (in 1988), this dataset became the National Minimum Data Set. Its predecessor is now referred to as the Historical Mortality Data Set (HMDS).

3.2.1.4 Statistics New Zealand Vitals File

In addition to the three NZHIS datasets, a fourth database contributed information to the NZCMS mortality file: the SNZ Vitals File. This file is SNZ's own file of births and deaths. When SNZ enter data from the death registration form (BDM28) they keep this information in their own Vitals file as well as passing it on to NZHIS. Thus, the SNZ Vitals file contains the same mortality information as the NMDS (since both are derived from the death registration form).

Since 1990, SNZ has assigned each decedent a residential meshblock code as well as the area unit code (which NMDS then converts to a domicile code). *This means that, for all individuals who died after 1990, the SNZ Vitals file includes a meshblock.* It was highly desirable to include a residential meshblock for each mortality record (as this provides much greater discriminatory power than a domicile code for record linkage). In constructing the mortality file for the NZCMS, we were able to retrieve meshblocks from the SNZ Vitals file for all those individuals who had died after 1990. This was simply achieved by using the registration office number, year of registration and registration number to create a unique identifier for each mortality record common to the NZHIS-provided mortality data and the SNZ Vitals File.

Figure 6: Sources of Data for the NZCMS Mortality File



From the above it can be seen that the sources of mortality data differed somewhat for each census cohort, depending on which databases existed at the time an individual died. This meant that slightly different variables were used in each record linkage, which in turn influenced the record linkage process itself. The details of the differences for each census cohort are discussed in more detail below.

3.2.2 Variables used in Mortality Records

Mortality records for the 1981 census cohort were derived entirely from the Historical Mortality Data Set (HMDS). Mortality records for the 1986, 1991 and 1996 census cohorts contained information from both the NHI and the NMDS databases. NZHIS linked the two databases using the NHI number. (The encrypted NHI number was retained on the mortality files at NZHIS on CD Rom, but not transferred to Statistics New Zealand in order to protect privacy). In addition, meshblock codes were added to mortality records for the 1991 and 1996 census cohorts from the SNZ Vitals file (as described above).

Census records used in the NZCMS were derived solely from the SNZ Census database, and thus contained only a single value for each variable. In contrast, many of the mortality records were derived from two independent databases (plus the meshblock code from the SNZ Vitals database for the 1991 and 1996 mortality records). This has several important implications:

1. **Mortality records contain two values for some variables (i.e. date of birth, sex, ethnic group).** This occurs where variables are recorded independently on both the NHI database and the NMDS health event database. Such variables may differ between the two databases for the same individual (e.g. changing self-defined ethnic group over time, coding errors for date of birth). Both values are included for record linkage, since we cannot tell which one will agree with the corresponding census record for the same person. Using both values increases the likelihood that we will find the true link for that mortality record.

2. **Several different domicile codes are included in each mortality record.** Domicile codes are the NZHIS equivalent of SNZ census area units. Several domicile codes (or area units) are available for any one individual: one from their NHI record, and one from *each* of their NMDS health event records. The NHI domicile code will be that entered directly by a hospital at a person's last health event. NMDS domicile codes correspond to a person's address at the time of various health events.

   We wanted a person's mortality record to include the domicile code that corresponded to their usual residence on the night of the census (in order to facilitate linkage with their census record). Domicile codes included in the mortality record comprise one from the NHI database (NHI-AU), and as many as three from the NMDS database. These latter three possibilities were: the code for the most recent health event prior to census night (pre-AU); the code for the first health event that occurred after census night (post-AU); and the code from their death registration form. (In practice, the death-event domicile code was derived from the meshblock recorded on the SNZ Vitals file, and is therefore designated 'Vitals-AU'.) These codes were selected in order to bracket as closely as possible the decedent's usual residence on census night.

Table 4: Mortality variables used in the record linkage

Variable	*Census-mortality cohort*	*Purpose*	*Comments*
Date of Birth (NMDS or HMDS)	1981 1986 1991 1996	•matching variable	The NMDS Death Event date of birth was used to generate the age at census night (in years), and the age for all decedents in the analysis of bias (i.e. differences between mortality records linked and those not linked).
Date of Birth (NHI)	1986 1991 1996	•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.
Sex (NMDS or HMDS)	1981 1986 1991 1996	•matching variable	-
Sex (NHI)	1986 1991 1996	•matching variable	As for date of birth, sex was available from both the NHI and NMDS Death Event File, usually independently sourced.
Ethnic Group (NMDS or HMDS)	1981 1986 1991 1996	•matching variable	For the 1981-1991 cohorts, the NMDS ethnic group was classified only as Maori, Pacific, and the Rest (the remainder), therefore it was specified as a matching variable with three values. For the 1996 cohort, ethnic group was classified as Maori, Pacific, Asian and non-Maori non-Pacific non-Asian.
Ethnic Group (NHI)	1986 1991 1996	•matching variable	As for date of birth, ethnic group was available from both the NHI and NMDS Death Event File, usually independently sourced. The NHI ethnicity was classified into five hierarchical levels: Maori, Pacific, Asian, Other, And European. However, it was aggregated for the actual linkage.
Country of Birth (NMDS or HMDS)	1981 1986 1991 1996	•matching variable	-
Meshblock •HMDS/NMDS •SNZ Vitals file	1981 1986 1991 1996	•blocking variable	The meshblock code is not available directly from NZHIS, but was merged with the mortality data from the SNZ Vitals file for 1991-94 and 1996-99. For the 1981-84 and 1986-89 mortality data, we had to geocode the addresses to 1996-base meshblocks (see text).
Vitals-AU (SNZ Vitals file)	1991 1996	•blocking variable	For most mortality records, simply the aggregate of the SNZ Vitals File meshblock. However, about 10% of mortality records had no meshblock on the SNZ Vitals File, but did have an area unit.

Variable	*Census-mortality cohort*	*Purpose*	*Comments*
Post-AU (NMDS Health Event)	1991 1996	•blocking variable	For many decedents it was possible to select a record for a health event (usually a hospitalisation) immediately after the relevant census, but before the death event. If the decedent had changed their usual residence between census night and death, the post-AU variable might give their actual area unit on census night, enabling a correct link between the decedent's mortality and census records.
Pre-AU (NMDS Health Event)	1991 1996	•blocking variable	As with the post-AU variable, it was possible for many decedents to select the health event record immediately before the relevant census. If the decedent had changed their usual residence between census night and death, the pre-AU variable might give their actual residential area unit on census night, enabling a correct link between the decedents mortality and census records.
NHI-AU (NHI file)	1986 1991 1996	•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-AU variables. Therefore, when a mortality record fails to link with a census record using other blocking variables, it may be that the NHI area unit as a blocking variable gives the correct usual residence for the decedent on census night, enabling a link with a census record.

3.2.3 Notes on Specific Variables

3.2.3.1 Ethnic Group

Mortality ethnicity: 1981, 1986 & 1991
Prior to 1995, ethnicity was recorded as a single option both on death forms and in the NHI dataset. This meant that for the 1981, 1986 and 1991 census cohorts, all mortality records used sole ethnic group as a matching variable.

For the 1981 census cohort ethnicity was available only from the HMDS dataset, so mortality records included only one value for ethnic group. For the 1986 and 1991 census cohorts, ethnicity data was available from both the NMDS and the NHI dataset (see Section 3.2.1). Mortality records for these cohorts therefore contain two values for ethnic group.

Up until July 1995, ethnic group was recorded on the NHI database as a single option.(Lewis 2002) NHI ethnicity is supposed to be self-defined – i.e. the hospital clerk is supposed to ask the patient their self-identified ethnic group. As a matching variable it was classified in five hierarchical levels: Maori, Pacific, Asian, Other, European. For the

purposes of record linkage, these levels were aggregated to three groups – i.e. Maori, Pacific and all other ethnicities (non-Maori non-Pacific).

Similarly, NMDS ethnicity was also recorded as a single option prior to September 1995. The NMDS ethnic group is collected by the undertaker, who is supposed to ask the family/whanau to 'self-identify' the decedent's ethnic group. In practice this often did not happen, and instead undertakers often selected what seemed to them the most likely ethnic group. NMDS ethnic group was classified only as Maori, Pacific, and non-Maori non-Pacific.

NHI is likely to be 'closer' to the census ethnic group (compared with NMDS ethnicity) as both are (in theory) self-defined.

Mortality ethnicity: 1996

- After 1995, both NMDS and NHI datasets changed to allow multiple options for an individual's ethnicity (as is the case for census data). This meant there now could be several values present for the ethnicity variable in both census and mortality records. As with the 1986 and 1991 census cohorts, mortality records contained two ethnicity variables – one from the NHI dataset and one from the NMDS. However, the 'sole' ethnicity options on the mortality data were not comparable to census 'sole' ethnicity. Therefore, the mortality ethnicity variable was specified as one composite variable as follows: 'Maori' if one of the ethnic groups on either the NMDS or NHI file was Maori; and of those remaining:
- 'Pacific' if one of the ethnic groups on either the NMDS or NHI file was Pacific; and of those remaining:
- 'Asian' if one of the ethnic groups on either the NMDS or NHI file was Asian; and the remaining values were classed as:
- 'non-Maori non-Pacific non-Asian'.

3.2.3.2 Geocodes

As described in section 3.1.2.2 above, geographical areas for census data can be categorised in different ways. All census records have two geographical codes: a meshblock and an area unit. In contrast, NZHIS records (including mortality records) have only a single geocode routinely recorded, called a ***domicile code.***

3.2.3.2.1 Forward Coding of Area Units

The NZHIS domicile code corresponds to one census area unit, but has a different numerical coding system. This is due to historical limitations in the size of the NZHIS database. The area unit code uses six digits, whereas the NZHIS domicile code uses only four digits; this four-digit code is derived from the area unit code assigned to each death record by SNZ. Thus at every point in time there is a one-to-one concordance between area unit and domicile codes (although forward coding to the 1996 base was no easy task!).

In order to use area units as a matching variable, it was necessary to have the same code used on both census and mortality records. This meant that, for some census cohorts, the

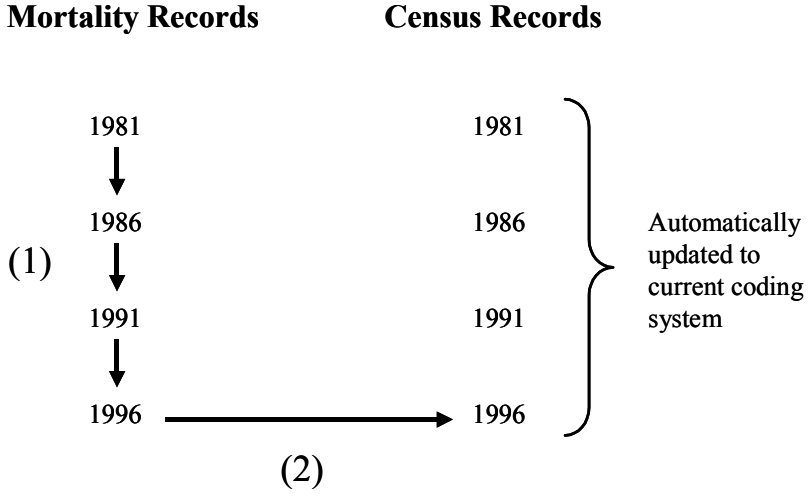
four digit NZHIS domicile codes had to be converted back to their original six digit area unit codes before the mortality records could be submitted for record linkage.

To further complicate matters, many area unit codes also had to be updated or 'forward coded'. As described above (in Section 3.1.2.2), geocodes have changed over time so that they represent a fairly standard-sized population unit. SNZ automatically updates geocodes on their census files, so census meshblocks and area units are always given in the most up-to-date coding system. At the time of the NZCMS, census cohorts from 1981, 1986 and 1996 used the 1996-base geocodes, while the 1991 census used the 1991-base geocode.

In contrast, the domicile codes used in the HMDS and NMDS databases are not updated, so records for individual health events use whichever coding system was current at the time the record was made. This meant that the domicile codes on the mortality records had to be updated to the 1996-base before they could be used for record linkage.

Forward coding of mortality geocodes thus involved two steps: (1) updating domicile codes to the 1996 domicile coding base; and (2) converting these 1996-base domicile codes to the equivalent 1996-base census area unit code.

Figure 7: Forward coding of mortality geocodes



Inaccuracies can creep in when codes do not have a straight one to one matching between years. Some old codes were split into two new codes (usually as a geographical area became more populated). Rarely, more than one old code was condensed into a single new code (usually when the population declined in an area). Occasionally part of an old code was recoded into one new value, while the rest was recoded into another new value.

3.2.4 Creation of Domicile Code Probe

Some of the domicile code data from the various sources was recorded using 1986 base codes, some used 1991 base codes and some used 1996 base codes. For some data we knew which year base was used, for some other data we were unsure and therefore had to develop a method of checking the data and finding out which base it was. We created a probe for each of 1981, 1986 and 1991 by developing a list of domicile codes that were unique to each year. These probes were then applied to the mortality dataset to determine which base year had been employed. We also made a list of unique domicile codes between 1991 and 1996 and used that to further separate the 1991- and 1996-coded data.

3.2.4.1.1 Assigning Meshblocks to Mortality Records

We also wanted to include a meshblock code for as many mortality records as possible. Meshblocks are included in all census records, but are not used in either HMDS/NMDS or NHI records. This meant we had to find alternative ways of assigning a meshblock to each mortality record.

For the 1991 and 1996 census cohorts, obtaining meshblocks for mortality records was reasonably straightforward as these were recorded from the SNZ Vitals file (see Section 3.2.1). SNZ simply inserted meshblocks directly from their Vitals file into the NMDS file from NZHIS.

For the 1981 and 1986 census cohorts, the only way to obtain meshblocks was to actually geocode a meshblock for each mortality record from the individual street address (recorded in the HMDS/NMDS and the NHI files).

The process of geocoding an address-derived meshblock code to over 80,000 mortality records was undertaken by three different providers: Datamail, Statistics New Zealand and New Zealand Health Information Services.

3.2.4.2 Datamail: 1986-89

Datamail is a company that specialises in geocoding. We provided them with a file consisting of an address and unique ID number for each mortality record in the 1986 mortality cohort. Datamail processed this file through their automatic geocoding system, assigning 1996-base meshblocks to as many records as possible.

3.2.4.3 SNZ Christchurch: 1986-89

Statistics New Zealand staff in Christchurch were then sent a file of the remaining mortality records that Datamail had been unable to geocode. SNZ assigned a meshblock to as many of these records as possible. Where they were unable to assign a meshblock, they were sometimes able to assign an area unit.

3.2.4.4 New Zealand Health Information Service: 1986-89

As a separate process, NZHIS also supplied meshblock and area unit codes for the majority of mortality records, using raw text address data. This was undertaken as part of a validation exercise for their newly established geocoding service.

3.2.4.5 New Zealand Health Information Service: 1981-84

Since the NZHIS geocoding service gave as good a return as Datamail, it was decided to rely solely on NZHIS in obtaining meshblocks for the 1981 mortality cohort. NZHIS duly undertook this process, and were able to assign meshblocks to a majority of mortality records. The remaining mortality records (which NZHIS had been unable to geocode) were then forwarded to SNZ Christchurch. SNZ assigned a meshblock or, as a second choice, an area unit to as many records as possible. Out of the 5,843 records submitted to them, SNZ assigned a meshblock to 2,808 and an AU to 2,214, leaving only 821 records for which geocodes could not be assigned.

3.2.5 Records Excluded from the Mortality File

3.2.5.1 Duplicate Records

When NZHIS was preparing the mortality data for this study they found some mortality records that appeared to be duplicates. Further clerical review of these 'possible duplicates' revealed a mixture of probable duplicates and probable non-duplicates.

It was clearly desirable to remove as many duplicates as possible before submitting the mortality file to record linkage. A standardised process was therefore developed for review of possible duplicate records (see 5.4.1 in Appendix). All possible duplicates identified by NZHIS were reviewed in this way, and probable duplicate records were amalgamated into a single record before progressing to record linkage.

3.2.5.2 Overseas Residents

Mortality records were excluded from the mortality file if the decedent was identified as an overseas resident (domicile code 9999). This was based on the assumption that most overseas residents would not have been in New Zealand on census night, and thus would not have a corresponding record in the census file. Also, the target population for the NZCMS is all New Zealand residents, so overseas residents were not included in the final cohort analysis.

3.2.5.3 Duration of Residence in New Zealand

Ideally, we would also like to exclude any New Zealand resident who died in New Zealand but had not been living in the country at the time of the previous census (since their mortality record would have no corresponding census record and thus could not be linked).

The NMDS mortality database includes a 'duration in New Zealand' variable, recording how long each decedent has been living in New Zealand. However, we found that for the majority of mortality records this variable was filled in with what seemed to be age at death. Because of this ambiguity, we created 6 groupings:

- 0 = Data where this field is missing or '99'. As the question on the death registration form only sought information if the deceased had ever lived overseas, these decedents were assumed to have been in New Zealand on census night.
- 1 = Data where the person's age is the same as this duration in NZ field. The majority of the data is in this group.
- 2 = People whose duration in New Zealand field does not equal their age, but from their date of death we can conclude that they were in NZ on census night.
- 3 = These are children under (or exactly equal to) 3 years of age, whose duration value was less than their age.
- 4 = These are decedents who have a value of zero for their duration in New Zealand field.
- 5 = Decedents who were not in NZ on census night. Their duration in NZ value is less than the time between their death and census night, the value is not blank or 99 or zero, nor is it equal to their age.

Following the record linkage, very few decedents with values of 3, 4 or 5 above were linked, confirming that they were probably not in New Zealand on census night. We therefore unlinked these records and removed all the relevant mortality records.

3.3. Interaction of country of birth and ethnicity

During the specification of the match run strategy, a problem became evident for links where there was either:

- agreement on ethnicity as Pacific and agreement on country of birth as Pacific; or
- agreement on ethnicity as Asian and agreement on country of birth as Asian.

Both these agreements attracted a high agreement weight due to low *u* probabilities. It was evident that many links with these two variable agreements were scoring a high total weight despite bad disagreements on DOB and sex. That is, many links with agreement on both Pacific (Asian) ethnicity and Pacific (Asian) country of birth that would definitely have been rejected on a clerical review were scoring a high total weight on probabilistic linkage. The reason was that decedents of Pacific or Asian ancestry were highly likely to have Pacific or Asian country of birth (respectively). Thus, we were in essence double counting an agreement.

To get around this problem, we created two additional variables for the linkage:

- EB_PAC coded as 5 on mortality data if both the mortality country of birth and ethnicity variable were Pacific, and coded as 2 on census data if both the census country of birth and ethnicity variable were Pacific. In all other instances the value was coded as missing.
- EB_ASIAN coded as 5 on mortality data if both the mortality country of birth and ethnicity variable were Asian, and coded as 2 on census data if both the census country of birth and ethnicity variable were Asian. In all other instances the value was coded as missing.

This 'trick' meant those links with complete agreement on ethnicity and country of birth when both variables were either Pacific or Asian actually registered as a disagreement for this particular composite variable (EB_PAC or EB_ASIAN) in Automatch®. (All other comparisons would have a missing value on either the census or mortality file, thereby scoring no weight.) For those with 'disagreement' on EB_PAC or EB_ASIAN we set the *m* probabilities at a value that gave a disagreement weight (in absolute terms) that was about the average of the agreement weight for Pacific/Asian ethnicity and country of birth. Therefore, the total weight for such comparisons was reduced to an amount that would be equivalent to only having one of the two variables in the matching process. Put in summary terms, we 'tricked' Automatch® to undo the double counting for Pacific and Asian people on two variables that were so highly correlated they violated the assumptions of probabilistic record linkage. We used this trick for the 1981 and 1996 census linkages.

3.4. Creating Blocking Variables

Each mortality record was assigned to a maximum of 16 blocks. (The maximum number of passes you can do in one linkage run is eight; thus this arrangement would allow two runs if necessary.) We used a set priority order to assign geocodes to the blocking variables 1 to 16 (i.e. variables BLOCK1 to BLOCK16). BLOCK1 was always a meshblock variable, as meshblock was the most discriminating blocking variable for the record linkage. We had two sources of meshblocks for 1986-89 mortality data, so BLOCK2 was also a meshblock value in this instance (the NZHIS meshblock value if different from the Datamail or SNZ meshblock value assigned as BLOCK1). BLOCK1 and BLOCK2 were left blank if no meshblock value was available.

The remaining blocking variables were assigned area unit values in 1996-base (except for the 1991 census cohort, which used 1991-base area units). At each step in the linkage process, an area unit value was assigned to a blocking variable only if it had not already been assigned to a higher priority order blocking variable. In this way, each blocking variable value was unique for each mortality record. If there were less than 16 unique possible area unit values for a given mortality record (true for the vast majority of mortality records), values were left blank.

The following sequence was used to prioritise the area unit value for 1986-89 mortality records:

a) Datamail/SNZ area unit after direct conversion from their geocoded meshblock value.
b) NZHIS area unit after direct conversion from their geocoded meshblock value.
c) First value on the list of all the NHI area unit values.
d) First value on the list of all the NMDS area unit values.
e) Second value on the list of all the NHI area unit values.
f) Second value on the list of all the NMDS area unit values.
g) Third value on the list of all the NHI area unit values.

etc until all the blocks were filled or the lists had all been used.

This same sequence was used for the 1981-84 mortality data, with the exception that there was only one meshblock code; consequently step b) was excluded, and each priority from c) down moved up one place.

These blocks are called BLOCK3 to BLOCK16, and we have stored the source of the data in BLKSRC3 to BLKSRC16 - e.g. NMDS2 means that it was the second item in the NMDS list.

The above process used our own modification to the concordance files that we obtained from NZHIS. The modifications involved using the distributions internal to our data of meshblock codes by domicile codes, which was demonstrably better than just using the original concordance files.

3.5. Non-geocode Variables

3.5.1 Country of Birth

The country of birth field in the NZHIS data used 1986 codes. In order to be linked with census data, this field was converted to 1996 codes, and grouped into nine categories (according to country/region). The converted values were then stored in the variable BP96 for 'birthplace 1996 code', and the aggregated data (grouped into nine categories) was stored in the variable called BIRTHCTY.

Chapter 4 Record linkage process and outputs

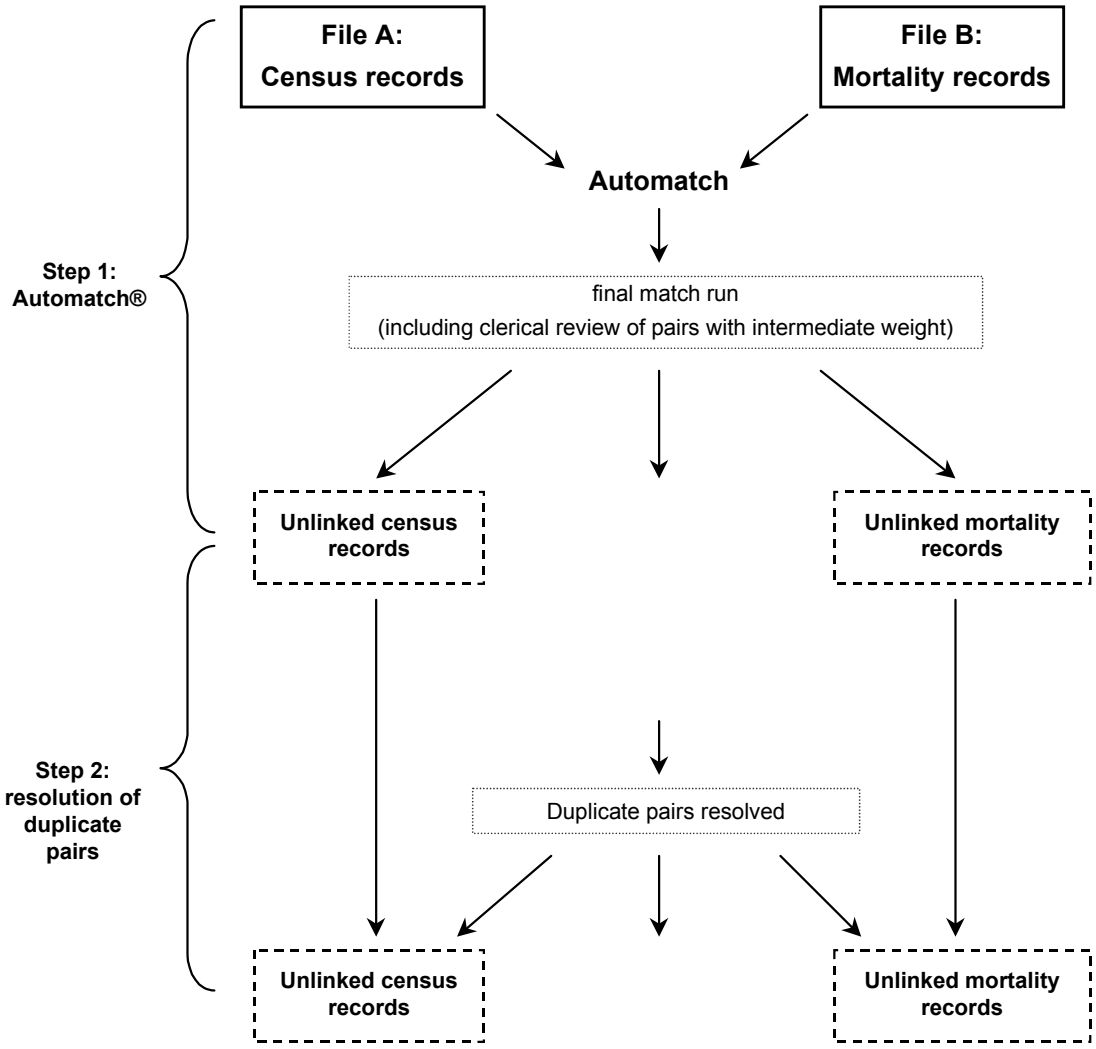
This chapter presents the outcomes of the record linkage process for each census year cohort, under the following headings:

1) data flow of mortality and census records
2) the final match-run strategy
3) the final m and u probabilities and variable weightings
4) accuracy of the record linkage (false positives and positive predictive value)

This ordering is intended to provide a logical outline of the linkage process, rather than reflecting the chronological sequence of the work. The order in which the linkage process for the four census cohorts was actually undertaken is as follows: 1991, 1986, 1996, and 1981. A more detailed description of the record linkage process and developmental work for 1991 (the first census to be linked) is presented here and elsewhere.(Blakely 2002)(Blakely 2001; Blakely et al. 2000) (The 1991 linkage was the first census linked to mortality records, and therefore subject to more developmental work and scrutiny.)

4.1. Overview of Linkage Process

Figure 8: Overview of linkage process



4.1.1 Step 1: Automatch® Linkage

As discussed in Chapter 3, mortality records were provided by New Zealand Health Information Services while census records came from the Statistics New Zealand database. Different approaches to record linkage were undertaken for each census cohort in order to find the strategy that produced the most accurate linkage results. Various match-run iterations were trailed by specifying different variables for blocking and matching, and varying the order in which these were used. Once the 'best' linkage strategy had been determined, this was used as the final Automatch® match-run. Clerical

review was included in the match-run wherever the 'cut-off' weight for a particular pass was specified as a range rather than a discrete value (see Section 2.1.3.)

The final match-run produced three output files:
- linked mortality and census records
- unlinked census records
- unlinked mortality records

These three Automatch® output files were extracted into SAS software for the next part of the linkage process.

4.1.2 Step 2: Resolution of Duplicate Pairs

For some mortality records there was more than one census record that produced a paired weight above the match cut-off. In this case Automatch® placed the matched pair with the highest weight (MP) in the linked file, while the 'extra' census record (DA) was kept 'associated' for later review. A smaller number of census records were paired with more than one mortality record; in this case the 'extra' mortality record (DB) was also kept associated for clerical review.

The next step in the linkage process was to resolve these duplicate pairs. Where one pair had a higher weight than the other pair, this first pair was accepted as a true link while the 'extra' record (from the second pair) was moved to the residual census or mortality file. Where duplicate pairs had equal weight, both pairs had to be discarded (since we were unable to determine which was the true matching pair). In this case all census components were returned to the residual census file, while all mortality components were moved to the residual mortality file.

4.1.3 Step 3: Removal of Ineligible Mortality Records

In all four census cohorts, the mortality file included records for people who had died in New Zealand but who had not been living in the country at the time of the previous census. Since these people had not filled in a census form, they did not form part of the census cohort and therefore should not have been included in the study. Also, since there was no corresponding census record, these mortality records could not (in theory) be linked.

The inclusion of these ineligible mortality records did not come to attention until after the linkage process had been carried out. It was therefore necessary to remove them from the Automatch® output files before the final cohort and bias datasets were formed. This removal comprises the final step in the linkage process.

4.1.4 Final Output Files

At the end of the linkage process we had three final output files: a final linked file, a final residual census file and a final residual mortality file (see Figure 8). From the numbers in

the linked and unlinked mortality files we can determine the percentage of eligible mortality records that were successfully linked for each census cohort.

Finally, we determined the accuracy of the record linkage process for each census cohort. This was achieved by estimating the number of false positive links and the positive predictive value for each pass in the final match-run. The weighted average of the pass-specific PPVs gives the overall PPV of the linkage process for that cohort.

NB: All census figures presented in this chapter have been 'random rounded' (RR) to protect confidentiality of census data.

4.2. 1981 linkage

4.2.1 Data flow of mortality and census records

4.2.1.1 Census and Mortality Files

The 1981 census included records for a total of 3,032,397(RR) New Zealand residents aged 74 or less on the night of the census: 24 March 1981 (File A). For the three years following the census, NZHIS received 44,932 mortality records for persons who had been 74 years or less on census night (File B). The flow of census and mortality records is shown in Figure 9.

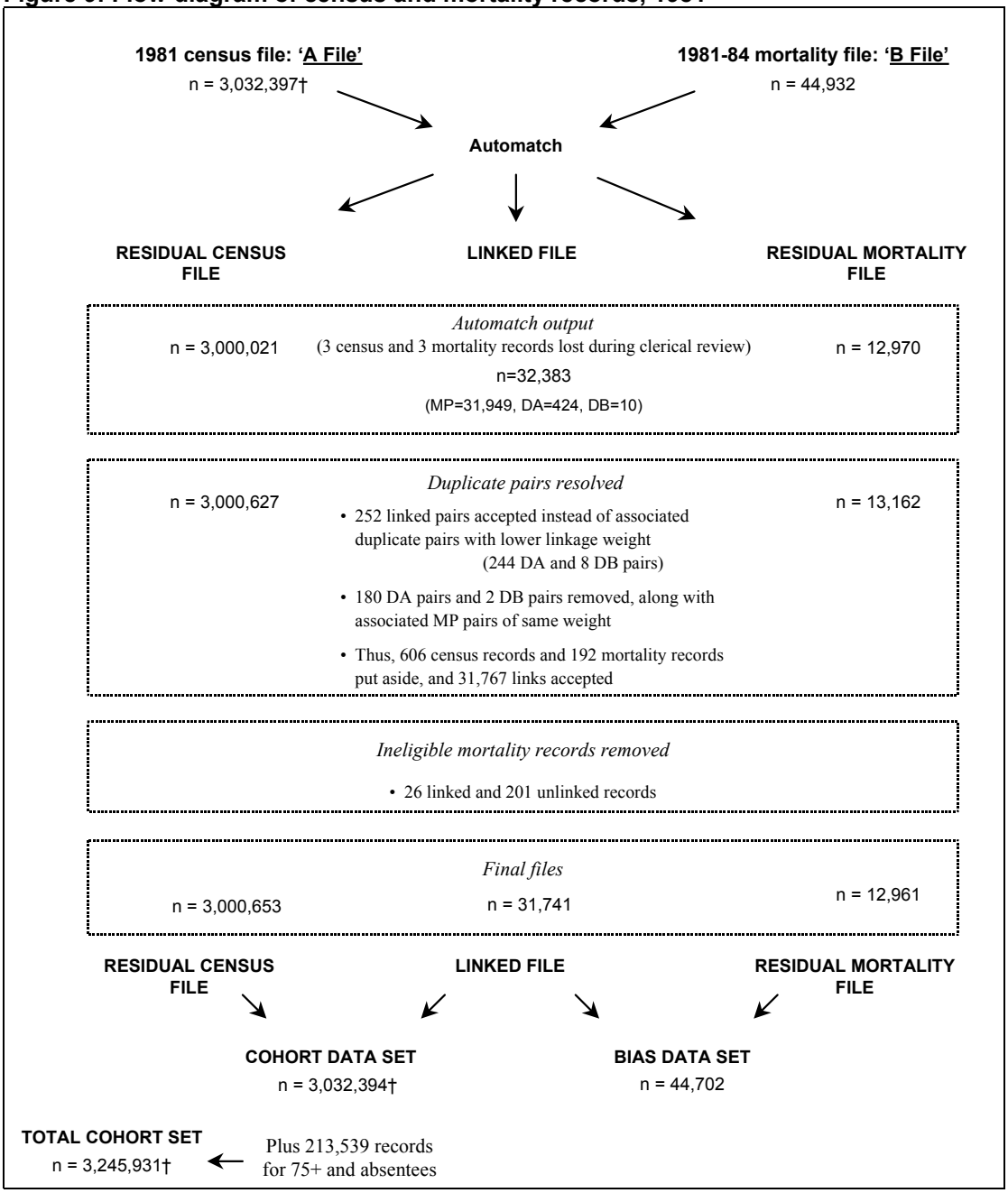
4.2.1.2 Automatch® Output

The final Automatch® linkage strategy produced 32,383 linked record pairs. During clerical review (of record pairs with an intermediate cut-off weight), three census and three mortality records were 'dropped' by Automatch®. The reason for this was not clear; however given the very small numbers involved and the time and resources required to repeat the final match run, it was decided to accept these six records as lost. The impact of such a small loss would have been negligible.

4.2.1.3 Resolution of Duplicate Pairs

252 MP pairs had a higher weight than their associated DA or DB pair(s), and were therefore retained as the 'best link'. 180 DA and 2 DB records were associated with MP pairs of equal linkage weight. Since there was no way of determining which of these pair combinations were the true links, the pairs were separated into their component census and mortality parts and these records returned to the residual census and mortality files.

Figure 9: Flow diagram of census and mortality records, 1981



† These numbers have been 'random rounded' to protect confidentiality of all census data

4.2.1.4 Ineligible Mortality Records

227 mortality records initially included in the mortality file were subsequently found to be ineligible, as the decedent had not been present at the time of the 1981 census. Of these ineligible records, 26 had been paired with a census record by Automatch®: these were deemed to be non-links (despite having reached the cut-off linkage weight). These pairs were separated into their component census records, which were returned to the residual census file, and the mortality records, which were removed altogether. The 201 ineligible mortality records that were unlinked were also removed.

4.2.2 Final match-run strategy

The final match run strategy is presented in Table 5. The majority of mortality record links (58.24%) were identified on the first pass, with a further 12.76% identified on passes 2 to 4. Overall, 71.01% of mortality records were linked to a census record.

Table 5: Final match-run strategy, 1981

Pass and blocking variable(s)	Main match specifications	Matching variables	Links (% of eligible mortality records)†	
1. Meshblock, Block 1	• Match cut-off weight 9.0 • +/- 1 tolerance for yyyy	• Sex, dd, mm, yyyy, ethnic group, birth country, EBPacCen‡: all from HMDS	26,036	(58.24%)
2. Area Unit, Block 2	• Match cut-off weight 10.0 • +/- 1 tolerance for yyyy	• Mm, sex, dd, yyyy, ethnic group, birth country, EBPacCen‡: all from HMDS	3,923	(8.78%)
3. Meshblock, Block 1	• Match cut-off weight 9.0 • +/- 1 tolerance for yyyy • Clerical review weight range 6.5 – 8.9	(As for Pass 1)	780	(1.74%)
4. Area Unit, Block 3	• Match cut-off weight 10.1 • +/- 1 tolerance for yyyy	(As for Pass 2)	1,002	(2.24%)
TOTAL			**31,741**	**(71.01%)**

4.2.3 Final *u* and *m* Probabilities

Table 6 lists the u and m probabilities for the final 1981 match-run.

Table 6: *u* and *m* probabilities, and agreement and disagreement weights for matching variables for the final match-run, 1981

Matching variable	Value	*m* probability	*u* probability	Agreement weight	Disagreement weight
Sex†	1 = Male	0.99	0.50	0.98	-6.37
	2 = Female	0.99	0.50	1.00	-6.39
Day of Birth †	Range	0.97	0.03	4.87 - 5.66	-4.99 - -5.04
Month of Birth †	Range	0.98	0.08 - 0.09	3.46 - 3.69	-5.50 - -5.53
Year of birth †	1910	0.99	0.01	7.11	-6.60
(examples by	1920	0.99	0.01	6.63	-6.59
decade)	1930	0.99	0.01	6.57	-6.59
	1940	0.99	0.01	6.30	-6.59
	1950	0.99	0.02	5.98	-6.59
	1960	0.99	0.02	5.67	-6.58
	1970	0.99	0.02	5.61	-6.58
	1980	0.99	0.02	5.92	-6.59
Ethnic group †	1 = Maori	0.80	0.13	2.67	-2.12
	2 = Pacific	0.80	0.03	4.65	-2.27
	3 = non-M, non-P	0.95	0.83	0.19	-1.74
Birthplace †	1 = NZ	0.95	0.86	0.14	-1.52
	2 = Australia	0.85	0.01	5.97	-2.71
	3 = British Isles	0.85	0.08	3.44	-2.62
	4 = Europe	0.85	0.01	5.96	-2.71
	5 = Pacific Is	0.85	0.02	5.48	-2.71
	6 = Africa	0.85	0.00	8.36	-2.71
	7 = Americas	0.85	0.00	7.52	-2.72
	8 = Asia	0.85	0.01	6.72	-2.74
	9 = Other	0.80	0.00	11.48	-2.32
Ethnicity / Birthplace Adjustment Factor	2 = PI born in Pacific (HMDS)	0.97	0.00	13.11	-5.06
	5 = PI born in Pacific (Census)	0.97	0.02	5.89	-5.04

† All from HMDS

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

Positive predictive values were estimated using the duplicate method. The PPV for passes 1 to 4 is shown in Table 7. The overall PPV for linkage of the 1981 cohort was estimated as 96.9%.

Table 7: Positive predictive value (PPV) and expected number of false positives (E[FP]) for passes 1 to 4 of the final match-run, 1981

Pass	Blocking Variable	Link pairs	Duplicate method E[FP]	PPV
1	Meshblock	26,036	158	99.4%
2	Area Unit	3,923	474	87.9%
3 + 4	Meshblock and Area Unit	1,782	356	80.0%
	Totals †	**31,741**	**988**	**96.9%**

4.3. 1986 linkage

4.3.1 Data flow of mortality and census records

4.3.1.1 Census and Mortality Files

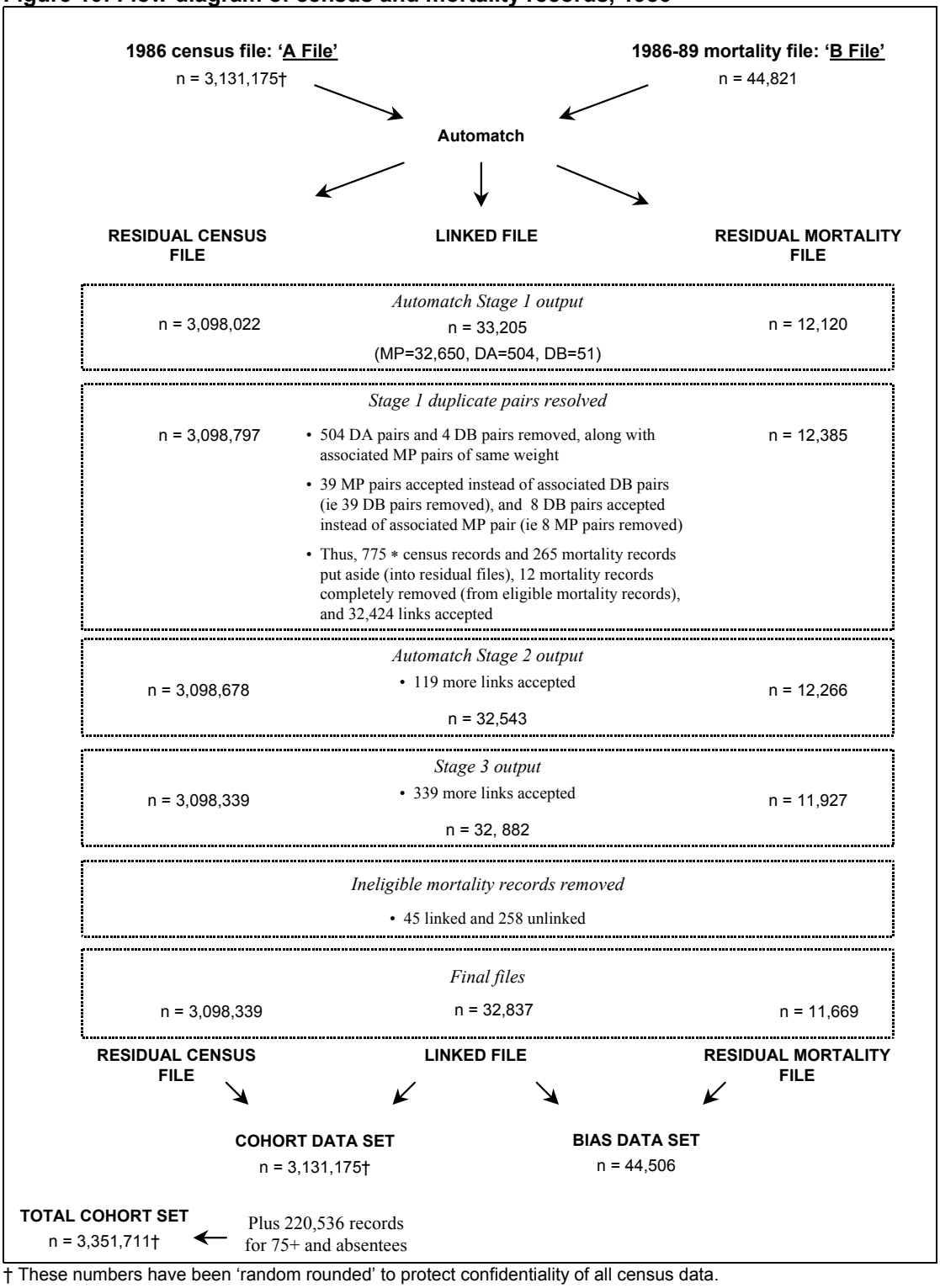
The 1986 census included records for 3,131,176(RR) New Zealand residents aged less than 75 years on the night of the census: 4 March 1986 (File A). For the three years following the census, NZHIS received 44,821 mortality records for New Zealand residents in this age cohort.

4.3.1.2 Automatch® Output

The Automatch® linkage process for the 1986 census cohort did not follow the usual pattern. This was because passes 1 to 5 were initially run without undertaking clerical review. Clerical review was usually conducted wherever the cut-off weight for a pass was given as a range, in which case all pairs with a linkage weight within that range were clerically reviewed before being accepted or rejected as links.

The linkage process for 1986 is therefore described in three stages. In Stage 1, passes 1 to 5 were undertaken without clerical review. In Stage 2, the same five passes were undertaken, but this time pairs with a linkage weight in the cut-off range were clerically reviewed before being accepted or rejected as links. In this way a further 119 links were obtained. Stage 3 was carried out using SAS software. For this part of the linkage process, pair combinations were accepted as a link if the census and mortality records had complete agreement for the variables of sex; day, month and year of birth; ethnic group; and birthplace. This stage allowed a further 339 links to be obtained.

Figure 10: Flow diagram of census and mortality records, 1986



† These numbers have been 'random rounded' to protect confidentiality of all census data.
∗ This figure has been derived and may be inaccurate.

Overall, this linkage strategy produced 32,837 linked pairs of census and mortality records (after resolution of duplicates and removal of ineligible records). At the end of Stage 1, there were 51 pairs with an associated duplicate mortality (DB) pair - a much higher number than for any of the other three cohorts. The reason for this was that the 1986 mortality file included a significant number of duplicated mortality records, a function of the implementation of the NHI number in 1988 and some duplication of

mortality records in the file presented to us by NZHIS. Thus the majority of DB pairs were due to the existence of two mortality records corresponding to the same census respondent.

4.3.1.3 Resolution of Duplicate Pairs

39 DB pairs were associated with an MP pair of higher linkage weight; in all these cases the MP pair was accepted in favour of the associated DB pair.

Because of the high number of duplicate mortality records, clerical review was undertaken for all DB pairs associated with an MP pair of equal weight. In some cases the DB pair was accepted as a link instead of its associated MP pair. This generally occurred where the two mortality records were deemed to apply to the same decedent, and the DB mortality record was considered to be more accurate than the MP mortality record (e.g. due to more accurate ICD coding). Thus eight DB pairs were accepted in favour of their associated MP pairs. The remaining four DB pairs were removed along with their associated MP pair – in these cases the two mortality records probably applied to different individuals (and were therefore 'genuine' DB/MP pairs rather than representing duplicated mortality records).

4.3.1.4 Ineligible Mortality Records

A total of 303 mortality records were found to be ineligible, as the decedent had not been in New Zealand at the time of the 1986 census. 45 of these mortality records were involved in links with census records; the 45 census records were returned to the residual census file, while the mortality records were removed altogether. A further 258 unlinked mortality records were also removed due to ineligibility.

4.3.2 Final match-run strategy

Table 8 presents the final match-run strategy and number of links by pass for the 1986 linkage process.

As described above, this linkage process took place in three stages. Passes 1 to 5 include both Stage 1 and Stage 2. In Stage 1, these five passes were undertaken using a discrete cut-off weight. In Stage 2 the same five passes were repeated, but this time using a cut-off range to select pairs that underwent clinical review. (The cut-off weight used in Stage 1 comprised the upper limit of the cut-off range used in Stage 2). For simplicity, these two stages are presented together in Table 8, along with the links obtained by each pass. These are the same results that would have occurred had passes 1 to 5 been conducted only once, using the clerical review range in the first place instead of a discrete cut-off weight.

Stage 3 of the process includes passes 6 and 7. These passes were actually carried out using SAS software, but performed the same function as passes conducted in Automatch®.

Table 8: Final match-run strategy, 1986

Pass and blocking variable(s)	Main match specifications	Matching variables	Links (% of eligible mortality records)[†]	
1. Meshblock: Block1	• Match cut-off weight 9.0	• Sex, dd, mm & yyyy and ethnic group from both NMDS and NHI • Birthplace from NMDS • dd, mm and yyyy prorated 1	26,126	(58.71%)
Stage 2	• Clerical review weight range 4.0-8.9	• dd, mm, yyyy all in agreement for NMDS and census		
2. Meshblock: Block2	• Match cut-off weight 10.0	(As for pass 1)	463	(1.04%)
Stage 2	• Clerical review weight range 4.0-9.9	(As for pass 1)		
3. Area Unit: Block3 and NMDS month of birth	• Match cut-off weight 11.0	• Sex, dd, yyyy and ethnic group from both NMDS/NHI • Birthplace from NMDS • dd & yyyy prorated 1	3,446	(7.74%)
Stage 2	• Clerical review weight range 4.0-10.9	• dd, mm, yyyy all in agreement for NMDS and census		
4. Area Unit: Block4 and NMDS month of birth	• Match cut-off weight 11.2	(Same as for pass 3)	1,560	(3.51%)
Stage 2	• Clerical review weight range 4.0-11.19	(Same as for pass 3)		
5. Area Unit: Block5 and NMDS month of birth	• Match cut-off weight 11.4	(Same as for pass 3)	903	(2.03%)
Stage 2	• Clerical review weight range 4.0-11.39	(Same as for pass 3)		
6. Stage 3	• SAS sweep programme	• Sex, dd, mm, yyyy, ethnic group, birthplace all in agreement on NMDS and census • Area unit equal to Block 6	220	(0.49%)
7. Stage 3	• SAS sweep programme	• Sex, dd, mm, yyyy, ethnic group, birthplace all in agreement on NMDS and census • Area unit equal to Block 7	119	(0.27%)
TOTAL			**32,837**	**(73.78%)**

4.3.2.1 Final *u* and *m* probabilities

Final u and m probabilities are presented in Table 9.

Table 9: *u* and *m* probabilities, and agreement and disagreement weights for matching variables for the final match-run, 1981

Matching variable	Value	m probability	u probability	Agreement weight	Disagreement weight
Sex †	1 = Male	0.99	0.50	0.97	-5.63
	2 = Female	0.99	0.50	0.99	-5.65
Day of Birth †	Range	0.97	0.03	4.86 to 5.65	-5.00 to -5.03
Month of Birth †	Range	0.98	0.08, 0.09	3.44 to 3.67	-5.50 to –5.52
Year of Birth (NMDS, examples by decade)	1920	0.98	0.01	6.69	-5.63
	1930	0.98	0.01	6.63	-5.62
	1940	0.98	0.01	6.37	-5.62
	1950	0.98	0.02	6.00	-5.50
	1960	0.98	0.02	5.80	-5.49
	1970	0.98	0.02	5.66	-5.61
	1980	0.98	0.02	5.95	-5.46
Ethnicity †	1 = Maori	0.80	0.09	3.08	-2.17
	2 = Pacific	0.80	0.03	4.65	-2.27
	3 = Non-M, non-P	0.95	0.87	0.12	-1.33
Birthplace (NMDS)	1 = NZ	0.99	0.85	0.22	-3.90
	2 = Australia	0.89	0.01	5.97	-3.18
	3 = British Isles	0.94	0.07	3.67	-3.96
	4 = Europe	0.94	0.01	6.01	-4.14
	5 = Pacific Is	0.90	0.02	5.28	-3.29
	6 = Africa	0.87	0.00	8.39	-2.95
	7 = Americas	0.86	0.01	7.39	-2.79
	8 = Asia	0.87	0.01	6.42	-2.90
	9 = Other	0.80	0.00	11.11	-2.32

† Both NMDS and NHI.

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

Table 10 presents the estimated positive predictive value for each of passes 1 to 5 and (combined) for passes 6 and 7. PPVs have been estimated by the duplicate method. They apply to the match-run strategy presented in Table 8 – i.e. as though Stages 1 and 2 of the linkage process had occurred concurrently.

The overall PPV for the 1986 linkage process was 96.7%.

Table 10: Positive predictive value (PPV) and expected number of false positives (E[FP]) for passes 1 to 7 of the final match-run, 1986

Pass	Blocking Variable	Link pairs	Duplicate method E[FP]	PPV
1	Meshblock	26,126	178	99.3%
2	Meshblock	463	20	95.7%
3	Area Unit	3,446	382	88.9%
4	Area Unit	1,560	291	81.4%
5	Area Unit	903	151	83.3%
6 + 7	(SAS sweep)	339	68	80.0%
	Totals †		**1090**	**96.7%**

4.4. 1991 linkage

4.4.1 Data flow of mortality and census records

4.4.1.1 Census and Mortality Files

The 1991 census included records for a total of 3,373,926(RR) New Zealand residents aged 74 or less on the night of the census: 5 March 1991 (File A). This is the number of New Zealand residents aged 74 years or less on census night (5 March 1991).

For the three years following the census, 42,229 mortality records were received from NZHIS. These records were then cross-linked with the SNZ Vitals file (in order to assign them with meshblocks). All but 46 mortality records were linked to a mortality record on the SNZ Vitals file, while 17 NZHIS mortality records were linked to two SNZ Vitals mortality records. Where this occurred we used both SNZ Vitals records plus the NZHIS record to 'create' two composite mortality records; thus we had 34 duplicate mortality records. All these records were submitted for record linkage, since we had no way of knowing which SNZ Vitals record gave the 'true' link.

Non-New Zealand residents were supposed to be excluded from the NZHIS mortality file on the basis of their domicile codes. In spite of this, the SNZ Vitals file provided a meshblock code of 'overseas usual residence' for 331 NZHIS mortality records. These records were therefore excluded from the mortality file. One of the 331 overseas residents was also one of the 34 duplicate mortality records 'created' from the original 17 NZHIS mortality records, so this record was removed. The 33 (34 minus 1) remaining duplicate mortality records were retained in the mortality file, in case the true link could be established later. (This was not possible, and all 33 duplicate records were eventually discarded). Thus a total of 41,915 mortality records (42,229 - 331 +17) were submitted to the record linkage (File B).

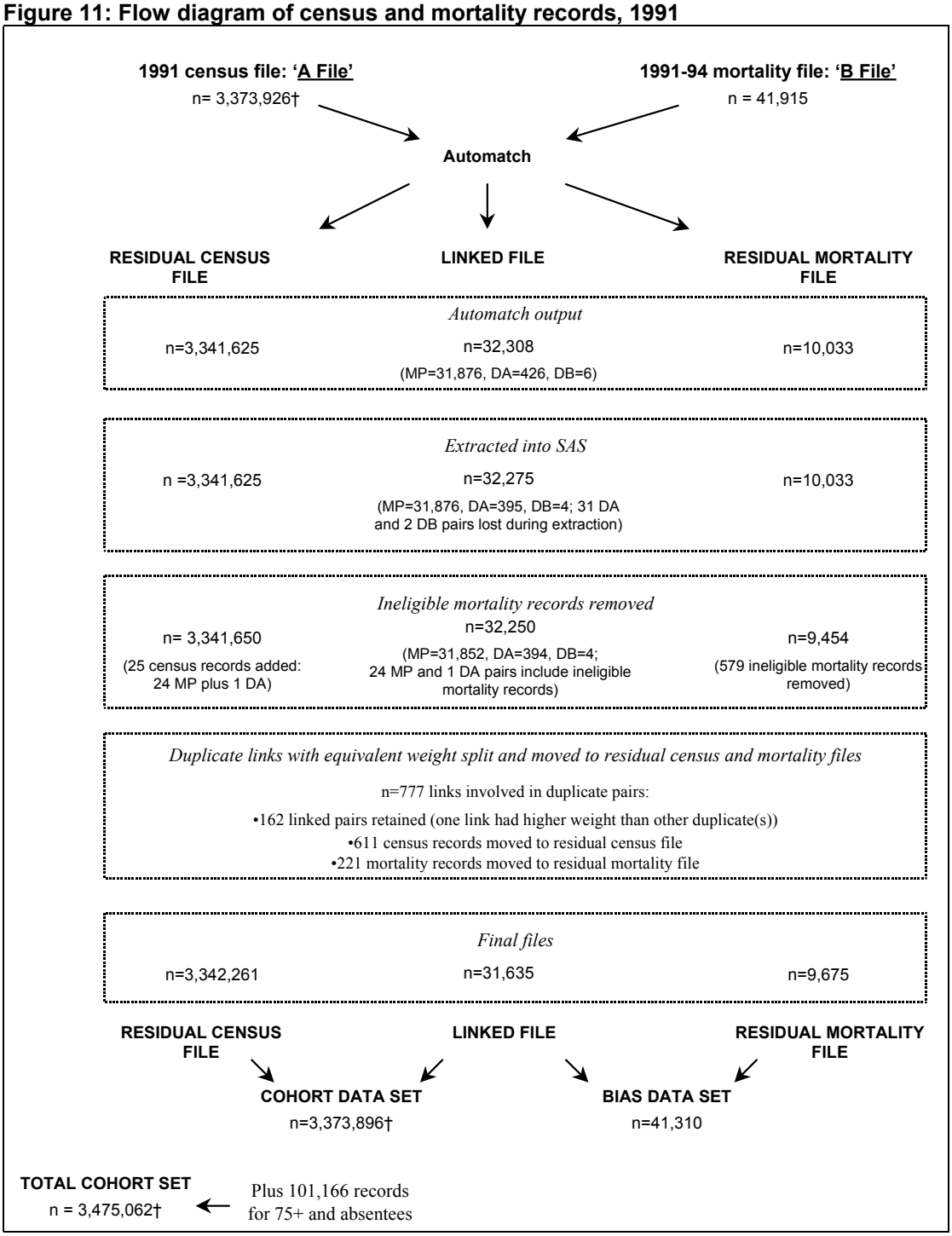
The flow of the mortality and census records is shown in Figure 11.

4.4.1.2 Automatch® Output

The final Automatch® linkage strategy produced 32,308 linked pairs of census and mortality records.

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 11: Flow diagram of census and mortality records, 1991



† These numbers have been 'random rounded' to protect confidentiality of all census data.

4.4.1.3 Ineligible Mortality Records

A number of mortality records originally included in the mortality file were subsequently deemed to be ineligible. These records were therefore removed before the output files were finalised.

The mortality data requested from 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 discrepancy was detected during the final match-run. Mortality records were also included for a further 38 decedents who died on the actual day of the census (5 March 1991). Further investigation suggested that the likelihood of these people completing a census (or having one completed on their behalf) was remote; these records were therefore also removed. Also included (knowingly) in the submitted mortality records were the 33 observations for the 17 NZHIS mortality records with two SNZ Vitals file links. Inspection of records suggested there would be little chance of successfully teasing apart the 17 duplicates, and it was decided to remove all 33 records.

Thus a total of 603 (532 + 38 + 33) '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 for repeating the final match-run of the record linkage.

Finally, a further data management issue requires stating for completeness. The census file counts in Figure 11 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 area unit 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.

4.4.1.4 Resolution of Duplicate Pairs

Altogether, 777 records were involved in an MP/DA or MP/DB duplicate association. 162 MP pairs had a higher match weight than their 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.

(NB: For other census cohorts, duplicate pairs were resolved before the final identification and removal of ineligible mortality records. The 1991 cohort was an exception to this pattern due to being the 'pilot study'.)

The final size of the linked file was 31,635. The sum of the linked file and residual census file records was 3,373,896, - i.e.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 – i.e. two less than the number of eligible mortality records (due to the loss of two DB pairs during extraction from Automatch®).

4.4.2 Final match-run strategy

The final match-run strategy, and number of links by pass, is presented in Table 11. 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. Brief details of each pass are given in the footnotes to Table 11.

Table 11: Final match-run strategy, 1991

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	(61.2 %)
2. Vitals-AU, 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-AU, and month of birth	(As for pass 2)	(As for pass 2)	1117	(2.70%)
4. Pre-AU, and month of birth	(As for pass 2)	(As for pass 2)	340	(0.82%)
5. NHI-AU, 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-AU 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%)**

4.4.2.1 Final *u* and *m* probabilities

Final *u* and *m* probabilities for the full match-run are shown in Table 12. The *m* probabilities are those determined by MPROB.

Table 12: *u* and *m* probabilities, and agreement and disagreement weights for matching variables for the final match-run, 1991

Matching variable	value	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 (NMDS, examples by decade)	1900	0.99	0.00	10.55	-6.14
	1910	0.99	0.00	8.24	-6.13
	1920	0.99	0.01	7.03	-6.24
	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 (NHI)	Maori	0.81	0.12	2.72	-2.22
	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 (NMDS)	Maori	0.73	0.12	2.56	-1.69
	Pacific	0.65	0.04	3.90	-1.45
	Rest	0.96	0.80	0.27	-2.53
Birthplace (NMDS)	NZ	0.99	0.80	0.31	-4.40
	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.

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

The estimated positive predictive values and number of false positives for the first five passes are shown in Table 13. Two methods were used to estimate the positive predictive

value for the 1991 cohort: the chance method and the duplicate method. These methods are described in detail elsewhere.(Blakely and Salmond in press)

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 (these are described elsewhere (Blakely et al. 1999)). If we assume the PPV to be 85% for the final three passes, 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 13: Positive predictive value (PPV) and expected number of false positives (E[FP]) for passes 1 to 5 of the final match-run, 1991

Pass	Blocking Variable	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-AU	3473	365	89.5%	274	92.1%
3	post-AU	1117	130	88.4%	134	88.0%
4	pre-AU	340	52	84.9%	39	88.5%
5	NHI-AU	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 (i.e. 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 matching record 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.

4.5. 1996 linkage

4.5.1 Data flow of mortality and census records

4.5.1.1 Census and Mortality Files

The 1996 census included records for 3,344,958(RR) New Zealand residents aged 74 years or less on census night: 5 March 1996 (File A). For the three years following the census, NZHIS received 39,665 mortality records for New Zealand residents in this age cohort.

4.5.1.2 Automatch Output

A large number of passes were used for linkage of the 1996 cohort. This was largely due to an inadvertent error in the blocking specifications for the first 15 passes. This error meant that many of these passes had to be repeated once the correct matching variable had been specified.

The above error occurred in specifying the blocking variable 'residence area unit' for census records. This variable was supposed to apply to a person's usual residence at the time of the census. Instead of using the usual residence, however, a specification for passes 1 to 15 referred to a person's address on the night of the census. For the majority of individuals, their address on census night and their usual residential address were the same: linkage for these records was unaffected by the specification error. However some individuals' usual address was different to their address on the night of the census. These census records could not be linked by passes 1 to 15, since they would have been allocated to a different block from any corresponding mortality record.

This mistake was discovered after the first 15 passes had been undertaken. Rather than beginning the linkage process over again, it was decided to select those census records that might have been affected by the specification error (i.e. those where the address on census night was different to the address of usual residence), and repeat the Automatch® passes using the correct variable for 'residence area unit'.

All those census records where 'address on census night' differed from 'address if usual residence' were therefore selected. Out of these selected records, 122 were already involved in links with mortality records. It was decided to un-pair these links and return the component records to the residual files: in this way any false links would be removed, while any true links would be re-linked in the repeated passes.

Eight of the first 15 passes were repeated (with the correct variable for ' residence area unit', as passes 16 to 23. These passes yielded a further 337 pairs which were accepted as links.

4.5.1.3 Resolution of Duplicate Pairs

Non-exact duplicate pairs were resolved after Pass 12, so that records that were not accepted as part of a linked pair could be submitted in subsequent passes. The remaining duplicate pairs were resolved after Pass 15. No further duplicate pairs occurred in passes 16 to 23 (which involved relatively low numbers of mortality records).

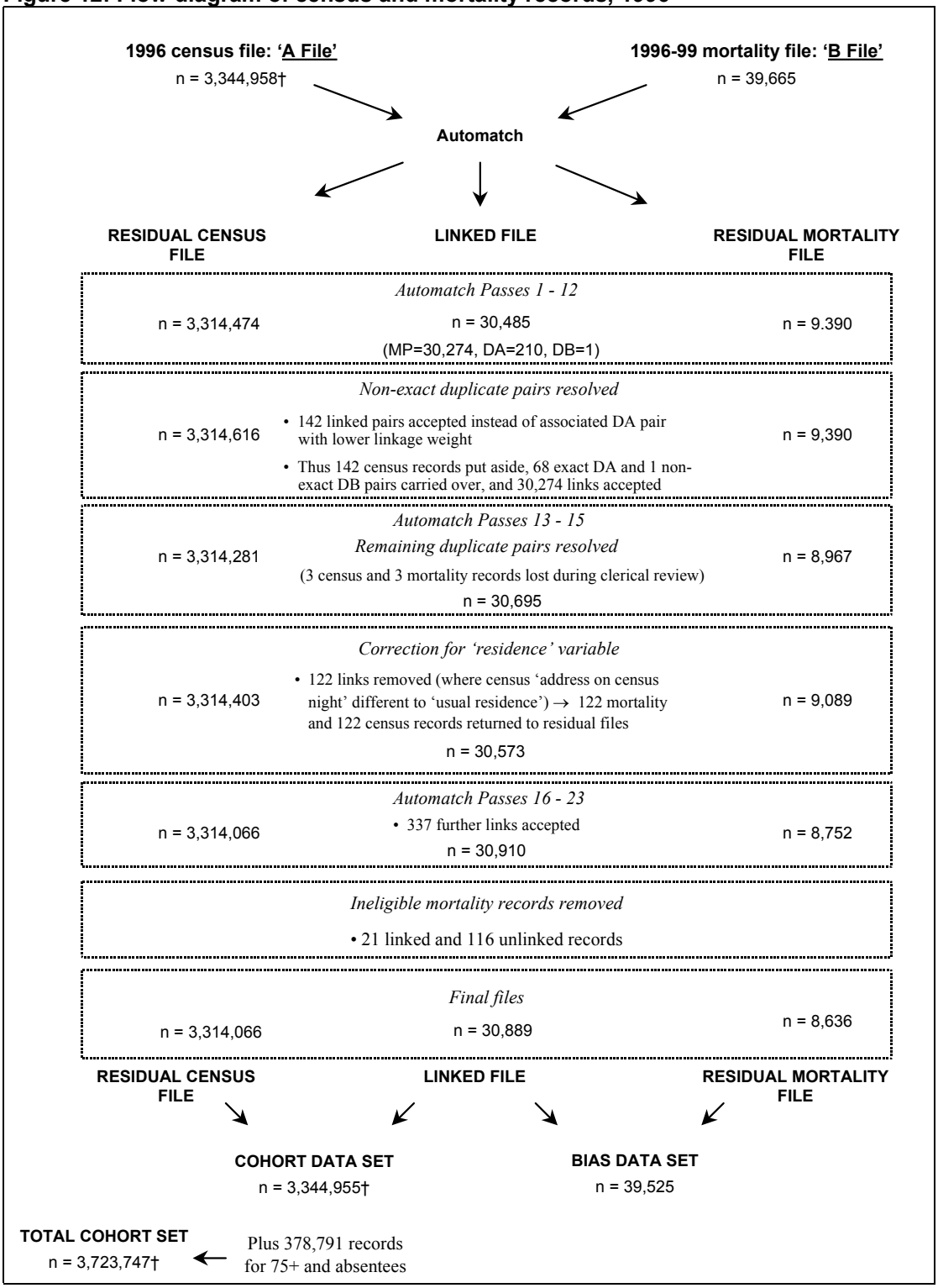
4.5.1.4 Ineligible Mortality Records

At the end of the Automatch® linkage process, 137 mortality records were found to be ineligible (as the decedent had not been resident in New Zealand at the time of the 1996 census). 21 of these records had been linked with a census record; the 21 census records in question were returned to the residual census file. All 137 ineligible mortality records were removed.

4.5.2 Final match-run strategy

The strategy for all 23 passes is shown in Table 14. The final linked file contained 30,889 pairs. In total, 78.15% of mortality records were linked with a census record.

Figure 12: Flow diagram of census and mortality records, 1996



† These numbers have been 'random rounded' to protect confidentiality of all census data

Table 14: Final match-run strategy, 1996

Pass and blocking variable(s)	Main match specifications	Matching variables	Links (% of eligible mortality records)[†]	
1. Meshblock: Block 1	• Match cut-off weight 11.0 • Clerical review weight range 8.5-10.9 • +/- 1 tolerance for yyyy	• Sex, dd, mm, yyyy, and ethnic group from both NMDS and NHI • Birthplace from NMDS • EBPacific, EBAsian from NMDS	24,239	(61.33%)
2. Census night area unit: Block 2; census month of birth, NMDS month of birth	• Match cut-off weight 12.0 • +/- 1 tolerance for yyyy	• Sex, dd, yyyy, and ethnic group from both NMDS and NHI • Birthplace from NMDS • EBPacific, EBAsian from NMDS	2,781	(7.04%)
3. Census night area unit: Block 3; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	2,247	(5.69%)
4. Census night area unit: Block 4; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	607	(1.54%)
5. Census night area unit: Block 5; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	97	(0.25%)
6. Census night area unit: Block 6; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	13	(0.03%)
7. Census night area unit: Block 7; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	4	(0.01%)
8. Census night area unit: Block 8; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	0	
9. Census night area unit: Block 2; census month of birth, NHI month of birth	(As for Pass 2)	(As for Pass 2)	24	(0.06%)
10. Census night area unit: Block 3; census month of birth, NHI month of birth	(As for Pass 2)	(As for Pass 2)	34	(0.09%)
11. Census night area unit: Block 4; census month of birth, NHI month of birth	(As for Pass 2)	(As for Pass 2)	11	(0.03%)
12. Census night area unit: Block 5; census month of birth, NHI month of birth	(As for Pass 2)	(As for Pass 2)	2	(0.01%)

Pass and blocking variable(s)	Main match specifications	Matching variables	Links (% of eligible mortality records)[†]	
13. Meshblock: Block 1 (As for Pass 1)	• Match cut-off weight 11.0 • Clerical review weight range 5.0 – 10.9 • +/- 1 tolerance for yyyy	(As for Pass 1)	424	(1.07%)
14. (As for Pass 2)	• Match cut-off weight 12.0 • Clerical review weight range 9.5 – 11.9 • +/- 1 tolerance for yyyy	(As for Pass 2)	69	(0.17%)
15. (As for Pass 3)	(As for Pass 14)	(As for Pass 2)	0	
16. Usual residence area unit: Block 2; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	161	(0.41%)
17. Usual residence area unit: Block 3; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	122	(0.31%)
18. Usual residence area unit: Block 4; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	45	(0.11%)
19. Usual residence area unit: Block 5; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	4	(0.01%)
20. Usual residence area unit: Block 6; census month of birth, NMDS month of birth	(As for Pass 2)	(As for Pass 2)	0	
21. Usual residence area unit: Block 2; census month of birth, NHI month of birth	(As for Pass 2)	(As for Pass 2)	0	
22. Usual residence area unit: Block 3; census month of birth, NHI month of birth	(As for Pass 2)	(As for Pass 2)	2	(0.01%)
23. Usual residence area unit: Block 4; census month of birth, NHI month of birth	(As for Pass 14)	(As for Pass 2)	3	(0.01%)
TOTAL			**30,889**	**(78.15%)**

4.5.2.1 Final u and m probabilities

Final *u* and *m* probabilities are presented in Table 15.

Table 15: *u* and *m* probabilities, and agreement and disagreement weights for matching variables for the final match-run, 1996

Matching variable	Value	*m* probability	*u* probability	Agreement weight	Disagreement weight
Sex †	1 = Male	0.99	0.50	0.98	-5.64
	2 = Female	0.99	0.50	0.98	-5.64
Day of Birth †	Range	0.97	0.03	4.81 – 5.68	-4.98 - -5.05
Month of Birth †	Range	0.98	0.08, 0.09	3.45 – 3.67	-5.50 - -5.54
Year of Birth †	1930	0.99	0.01	6.83	-6.59
	1940	0.99	0.01	6.54	-6.55
	1950	0.99	0.01	6.13	-6.63
	1960	0.99	0.02	5.85	-6.63
	1970	0.99	0.02	5.99	-6.09
	1980	0.98	0.02	6.02	-6.03
	1990	0.98	0.02	5.82	-5.29
Ethnic group ‡	1 = Maori	0.80	0.12	2.73	-2.13
	2 = Pacific	0.80	0.05	4.12	-2.25
	4 = Asian	0.80	0.05	4.11	-2.25
	5 = NonMaori, NonPacific, NonAsian	0.95	0.79	0.26	-2.08
Birthplace (NMDS and Census)	1 = NZ	0.99	0.81	0.28	-4.22
	2 = Australia	0.85	0.02	5.77	-2.72
	3 = British Isles	0.85	0.06	3.80	-2.64
	4 = Europe	0.85	0.02	5.74	-2.71
	5 = Pacific Is	0.85	0.03	4.85	-2.69
	6 = Africa	0.85	0.01	7.31	-2.76
	7 = Americas	0.85	0.01	7.03	-2.71
	8 = Asia	0.85	0.03	4.62	-2.68
	9 = Other	0.82	0.00	8.80	-2.47
Ethnicity / Birthplace Adjustment Factor	2 = PI, born in Pacific (NMDS)	0.92	0.00	11.05	-3.64
	5 = PI, born in Pacific (Census)	0.92	0.02	5.29	-3.62
Ethnicity / Birthplace Adjustment Factor	4 = Asian, born in Asia (NMDS)	0.92	0.00	12.61	-3.63
	8 = Asian, born in Asia (Census)	0.92	0.03	4.85	-3.62

† Both NMDS and NHI (as an array), and Census.

‡ Ethnicity from Mortality data (Eth_Match) and Census data (Sole and Prioritised).

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

Estimated positive predictive values for the Automatch® passes are presented in Table 16. The majority of links were obtained from passes 1 to 4, which have a high PPV. Passes 5 to 23 had a lower PPV, but contributed only a small number of links (i.e. 203).

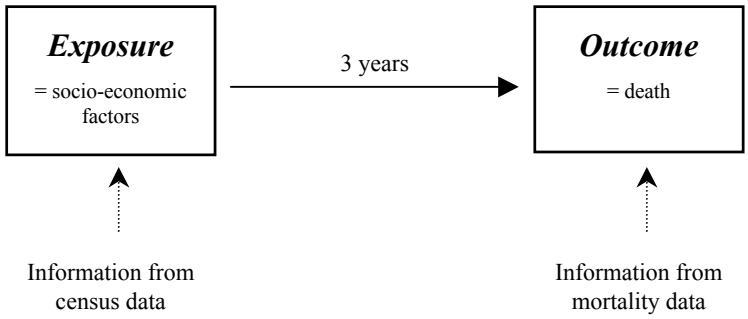
Overall, the PPV for the 1996 cohort was 97.4%.

Table 16: Positive predictive value (PPV) and expected number of false positives (E[FP]) for passes 1 to 23 of the final match-run, 1996

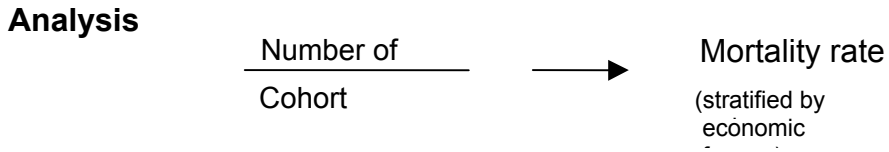
	Pass	Link pairs	Duplicate method E[FP]	PPV
1	Meshblock	24,239	84	99.7%
2	Area Unit	2,781	303	89.1%
3	Area Unit	2,247	173	92.3%
4	Area Unit	607	36	94.2%
5 - 23	Area Unit or Meshblock	1,015	203	80.0%
	Totals †	**30,889**	**799**	**97.4%**

Chapter 5 Cohort, bias and unlock files

The NZCMS is a cohort study, where the cohort consists of the entire population of New Zealand and the follow-up period is the three years after each census. The exposures of interest are socio-economic factors recorded in census data (although any one of the variables recorded on census forms could also be used as an 'exposure'). The outcome of interest is death in the three years following census night, for people aged 0 – 74 years on census night. The primary aim of the study is to determine mortality rates within different socio-economic strata of the New Zealand population, and thus estimate the association between socio-economic factors and mortality.

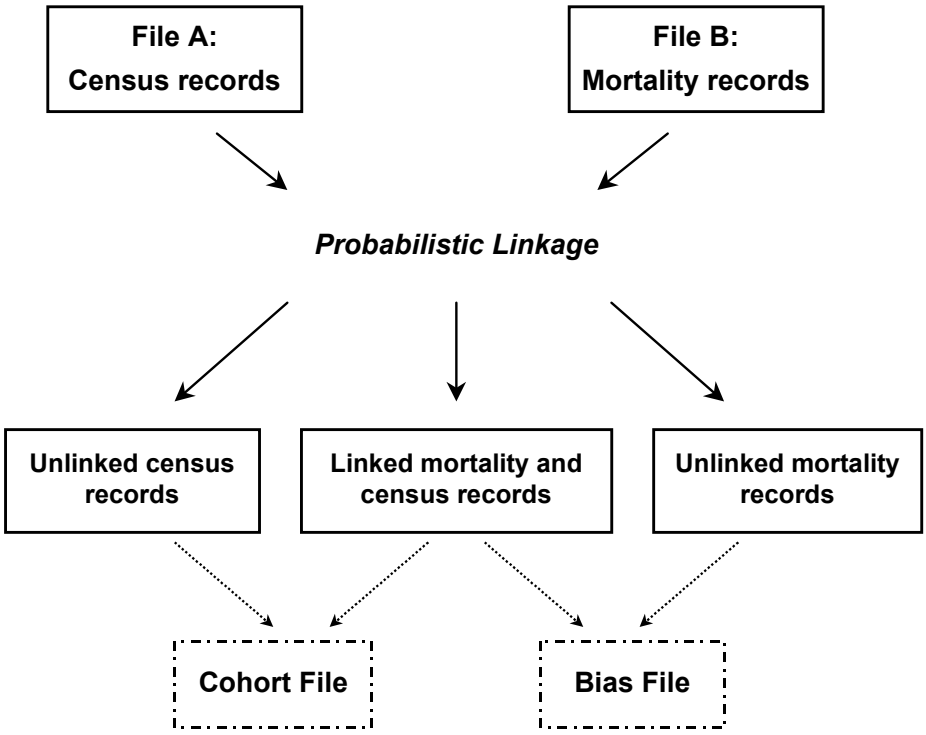


Chapter 2 to Chapter 4 of this report describe the methodology, data requirements, linkage process and outputs involved in the NZCMS. All these processes are required in order to prepare the data required for the cohort study. The actual point of the study – the data analysis – comes only at the end of this complex linkage process.



The linkage process produces three final files: the residual (unlinked) census records, the linked mortality-census records, and the residual (unlinked) mortality records. These three files form the basis of our data analysis.

There are three aspects to analysis of the NZCMS data:



5.1.1 Cohort Analysis

The cohort analysis is the primary objective of the NZCMS – i.e. determining mortality rates within different socio-economic strata of the New Zealand population. Mortality data (the numerator) is derived from linked mortality records, while census data (the denominator) is derived from both linked and unlinked census records.

Socio-economic 'exposures' derived from census data include small area deprivation, education, labour force status, car access, housing tenure, and household income. Covariates include age, sex, ethnicity, receipt of a sickness benefit, and marital status.

5.1.2 Bias Analysis

This refers to analysis of bias occurring in the linkage process – i.e. estimating systematic differences between linked and unlinked mortality records. Bias analysis involves comparison of mortality records in the linked census-mortality file with those in the residual (unlinked) mortality file. It is important to measure any bias in the linkage process, as this will cause the results of cohort analysis to be biased (since cohort analysis is based solely on linked mortality records).

5.1.3 Unlock

'Unlock' means measuring the extent to which historical Maori and Pacific mortality rates have been underestimated. This underestimation occurs because of the differences in classifying ethnicity on mortality and census records. Census ethnicity is derived from self-identified ethnic group (as recorded on the census form). Mortality ethnicity is derived from death records, which are completed by undertakers.

Many individuals whose self-identified (census) ethnic group was 'Maori' or 'Pacific' were categorised as 'non-Maori non-Pacific' on mortality records. This misclassification bias results in an underestimate of mortality rates in Maori and Pacific Islanders. This is because mortality data (the numerator) underestimated Maori and Pacific deaths relative to census data (the denominator), which records of self-identified ethnic group.

By measuring this numerator-denominator bias, we can calculate adjustment ratios that can be used to correct historical estimates of mortality in Maori and Pacific Islanders. We refer to this adjustment as 'unlocking the numerator-denominator bias'.

5.2. Variables included in the cohort file

Those variables included in the cohort file are presented in the following table. (Technical Reports 4 and 5 detail the Unlock and Bias data-sets more comprehensively.)

Table 17 lists each variable name (as recorded in Datalab); the census cohort(s) to which the variable applies and the variable format used for each cohort (as used in SAS); and the variable label, which generally describes the variable in question.

In most cases the variable label in Table 17 is self-explanatory. For other variables, longer explanatory notes are provided in the pages following the tables; these should be consulted where a more comprehensive description of a variable is required.

A detailed list of variable formats (as used in SAS) is included in the Appendix (page 96).

Table 17: Variables used in cohort analysis

Variable Name	Format 1981	Format 1986	Format 1991	Format 1996	Variable Label
AbsentFlg	fabs.	fabs.	fabs.	fabs.	Absentee Indicator
AgeC_5yr	f5AgG.	f5AgG.	f5AgG.	f5AgG.	Age at Census (5 year age groups)
AgeC_Gp	fAgeC.	fAgeC.	fAgeC.	fAgeC.	Age at Census (5 Std Groups)
AgeC_mths	f5AgM.	f5AgM.	f5AgM.	f5AgM.	Age at Census (months)
AgeC_yrs	f5AgY.	f5AgY.	f5AgY.	f5AgY.	Age at Census (years)
AgeD_mths	f5AgM.	f5AgM.	f5AgM.	f5AgM.	Age at Death (months)
AU1Yr	fYesNo.	-	-	-	Same Area Unit of Residence 1 Year Ago
BabyBrn	-	-	-	fBBrn.	Number of Live Babies Given Birth To
BirthGp	fbthgp.	fbthgp.	fbthgp.	fbthgp.	Country of Birth
CauseDeath	f4dth.	f4dth.	f4dth.	f4dth.	Cause of Death (4 groups)
CenYear	[1981]	[1986]	[1991]	[1996]	Year of Census
DisCode	-	-	-	fDisCd.	Long-Term Disability or Handicap
DisInd	-	-	-	fDisIn.	Disability Indicator (from HealthProb & DisCode)
EdLAllCur	fAtLev.	-	-	-	Current Education Attendance Level
EdLAllHgh	fEdAtt.	-	-	-	Highest Level of Education Attendance
EdLAllPst	fAtLev.	-	-	-	Past Education Attendance Level
EdLSchHgh	fscat.	-	-	-	School Attendance Level
EdQAll_A	f81qual.	-	-	-	First Grouped Qualification Gained
EdQAll_B	f81qual.	-	-	-	Second Grouped Qualification Gained
EdQAll_C	f81qual.	-	-	-	Third Grouped Qualification Gained
EdQAll_D	f81qual.	-	-	-	Fourth Grouped Qualification Gained
EdQAllHgh	f81HQal.	f86HQal.	f91HQ.	f96HQal.	Highest Qualification Obtained; Highest Qualification Gained (SNZ Protocol); Highest Qualification Obtained; Derived Highest Qualification Gained
EdQAllHghDet	-	-	-	f96HQ.	Highest Qualification Gained
EdQSchHgh	f81sql.	f86sql.	f91sql.	f96sql.	Highest School Qualification
EdQTer_A	-	-	f91TQa.	f96Ter.	Tertiary Qual Gained, Group A; Tertiary Qual 1 Attainment Level
EdQTer_B	-	-	f91TQb.	f96Ter.	Tertiary Qual Gained, Group B; Tertiary Qual 2 Attainment Level
EdQTer_C	-	-	f91TQc.	-	Tertiary Qual Gained, Group C
EdQTer_D	-	-	f91TQd.	-	Tertiary Qual Gained, Group D
EdQTer_E	-	-	f91TQe.	-	Tertiary Qual Gained, Group E

Variable Name	Format 1981	Format 1986	Format 1991	Format 1996	Variable Label
EdQTerHgh	f81TQ.	f86tql.	-	-	Tertiary Qualification Gained
EmpSt	f81Emp.	f86Emp.	-	f96Emp.	Employment Status
EthCenDet	-	-	f91EthD.	-	Ethnicity -Detailed
EthCenGp3	f4eth.	-	-	-	Ethnicity
EthCenGp4	f4eth.	-	-	-	Ethnicity
EthCenGp6_A	-	fdeth.	fdeth.	fdeth.	Ethnicity -A
EthCenGp6_B	-	fdeth.	fdeth.	fdeth.	Ethnicity -B
EthCenGp6_C	-	fdeth.	fdeth.	fdeth.	Ethnicity -C
EthCenPr3	-	f4eth.	f4eth.	f4eth.	Ethnicity -Prioritised
EthCenPr4	-	f4eth.	f4eth.	f4eth.	Ethnicity -Prioritised
EthCenPr5	-	-	fnhiraw.	fraw.	Ethnicity -Prioritised*
EthCenSol3	-	f4eth.	f4eth.	f4eth.	Ethnicity -Sole; Ethnicity -Sole*; Ethnicity -Sole
EthCenSol4	-	f4eth.	f4eth.	f4eth.	Ethnicity -Sole
EthCenSol5	-	-	fnhiraw.	-	Ethnicity -Sole*
FamCode	fFamC.	fFamC.	-	-	Family Code
FamType	-	-	f91FamT.	f96FamT.	Family Type
G_AHB	f89AHB.	f89AHB.	f89AHB.	f89AHB.	Area Health Board 1989
G_AHBD91	-	-	f91AHD.	-	Usual Residence Area Health Board Consituent District
G_AHD	f93AHD.	f93AHD.	f91AHD.	f93AHD.	Area Health District 1993
G_RHA	frha.	frha.	frha.	frha.	Regional Health Authority (1989 AHB)
G_Rurality	frural.	frural.	frural.	frural.	Rurality Indicator
G_TLA5yr	-	-	-	f95tla.	TLA 1995 Address 5 Years Ago
G_TLA89	-	-	f95tla.	-	Territorial Local Authority 1989
G_TLA95	f95tla.	f95tla.	-	f95tla.	Territorial Local Authority 1995
G_UA91	-	-	f91UA.	-	Usual Residence Urban Area 1991
G_UA96	f96UA.	f96UA.	-	f96UA.	Usual Residence Urban Area 1996
H_BCars	f8num.	-	-	-	Number of Business Cars in H/H
H_Bdrms	f20num.	f8num.	f8num.	f14num.	Number of Bedrooms
H_DwgTp	f81dtyp.	-	-	f96dtyp.	Dwelling Type (detailed); Dwelling Record Type
H_DwgTpG	-	fdtyp.	fdtyp.	fdtyp.	Dwelling Type
H_FtJob	-	f7num.	-	-	Number of Full-time Jobs in H/H
H_IncAC	-	-	-	fIncS.	H/H Inc. Srce - ACC Regular Payments
H_IncDP	-	-	-	fIncS.	H/H Inc. Srce - Domestic Purposes Benefit

Variable Name	Format 1981	Format 1986	Format 1991	Format 1996	Variable Label
H_IncGB	-	-	-	fIncS.	H/H Inc. Srce - Other Government Benefits
H_IncIB	-	-	-	fIncS.	H/H Inc. Srce - Invalids Benefit
H_IncSB	-	-	-	fIncS.	H/H Inc. Srce - Sickness Benefit
H_IncSE	-	-	-	fIncS.	H/H Inc. Srce - Self-employment
H_IncUB	-	-	-	fIncS.	H/H Inc. Srce - Unemployment Benefit
H_IncWS	-	-	-	fIncS.	H/H Inc. Srce - Wages/Salary etc.
H_Mveh	f8num.	f5num.	f5num.	f3num.	Number of Private Cars in H/H; Number of Motor Vehicles in H/H; Number of Motor Vehicles in H/H; Number of Motor Vehicles in H/H
H_NAbCh	f9num.	-	-	-	Number of Children Absent in H/H
H_NAbTot	f9num.	-	-	f5num.	Total Number of Absentees in H/H
H_NAdult	f8num.	f8num.	-	-	Number of Adults aged 20+ in H/H (on C/N); Number of Adults aged 16+ in H/H (on C/N)
H_NChn	f8num.	f8num.	-	-	Number of Children aged 0-15 in H/H (on C/N)
H_NOccy	f81Occ.	f81Occ.	f91Occ.	f96Occ.	Nature of Occupancy
H_OccTot	-	-	-	f500nm.	Total Number of Occupants in H/H
H_PBike	f8num.	-	-	-	Number of Pushbikes in H/H
H_PerFam	-	-	-	f20num.	Number of People in Family
H_PtJob	-	f7gnum.	-	-	Number of Part-time Jobs in H/H
H_Teleph	-	-	-	fTele.	Telephone in Dwelling
H_Tenure	-	-	-	f96Tenr.	Tenure
H_THInc	f81Inc.	f86Inc.	f91Inc.	f96Inc.	Total Household Income
H_Type	f81HHT.	-	-	-	Household Type
H_UsHHC	f81UHC.	fhhc.	-	fhhc.	Usual Household Composition
HealthProb	-	-	-	fHProb.	Health Problems
HealthProb_A	-	-	-	fHProbD.	Health Problem 1
HealthProb_B	-	-	-	fHProbD.	Health Problem 2
HealthProb_C	-	-	-	fHProbD.	Health Problem 3
HrsWk	f81hwk.	-	-	-	Total Hours Worked (per week)
HrsWkG	-	-	-	fhwk.	Total Number of Hours Worked
I_DPB	fiDPB.	fiDPB.	-	-	Domestic Purposes Benefit
I_FamBen	fiFB.	fiFB.	-	-	Family Benefit
I_FamCare	-	fiFC.	-	-	Family Care Benefit
I_IncSup	fiIS.	fiIS.	-	-	Income Support Payments Indicator
I_InvalBen	fiIB.	-	-	-	Invalids Benefit Indicator

Variable Name	Format 1981	Format 1986	Format 1991	Format 1996	Variable Label
I_ISP_Der	-	-	fi91ISP.	-	Income Support Payments -Derived
I_ISPA	-	-	fi91IG.	-	Income Support Payment Group A
I_ISPB	-	-	fi91IG.	-	Income Support Payment Group B
I_ISPC	-	-	fi91IG.	-	Income Support Payment Group C
I_ISPD	-	-	fi91IG.	-	Income Support Payment Group D
I_ISPE	-	-	fi91IG.	-	Income Support Payment Group E
I_OispG	-	-	fi91IO.	-	Other Income Support Payments -Grouped
I_PIS_AC	-	-	-	fIncS.	Personal Inc. Srce - ACC Regular Payments
I_PIS_DB	-	-	-	fIncS.	Personal Inc. Srce - Domestic Purposes Benefit
I_PIS_GB	-	-	-	fIncS.	Personal Inc. Srce - Other Government Benefit
I_PIS_IB	-	-	-	fIncS.	Personal Inc. Srce - Invalids Benefit
I_PIS_SB	-	-	-	fIncS.	Personal Inc. Srce - Sickness Benefit
I_PIS_SE	-	-	-	fIncS.	Personal Inc. Srce - Self-employment
I_PIS_UB	-	-	-	fIncS.	Personal Inc. Srce - Unemployment Benefit
I_PIS_WS	-	-	-	fIncS.	Personal Inc. Srce - Wages/Salary etc.
I_SickBen	fiSick.	fiSick.	-	-	Sickness Benefit
I_TInc	f81Inc.	f86Inc.	f91Inc.	f96Inc.	Total Personal Income
I_UnEmpBen	fiUB.	fiUB.	-	-	Unemployment Benefit
ICD_Gp	$ficd.	$ficd.	$ficd.	$ficd.	International Cause of Death (ICD)
ID_Cohort	[Cnnnnnnn]	[Cnnnnnnn]	[Cnnnnnnn]	[Cnnnnnnn]	Unique Cohort Id
ID_Dwell	[Dnnnnnnn]	[Dnnnnnnn]	[Dnnnnnnn]	[Dnnnnnnn]	Unique Dwelling Id
Imp	-	-	f91IImp.	-	Imputation Indicator
ImpAge	-	-	f91IAge.	f96IAge.	Age Imputation Indicator
ImpForm	-	-	-	f96IDum.	Form Imputated Indicator (Dummy Form)
ImpLFS	-	-	-	f96ILFS.	Imputation in Labour Force Status
ImpMonth	fIMth.	fIMth.	fIMth.	fIMth.	Month of Age Imputation Indicator
ImpRes	-	-	-	f96IRes.	Imputation in Usual Residence Status
ImpSex	-	-	-	f96ISex.	Imputation in Sex
IndAnz1	-	-	-	$fANZ.	ANZSIC Industry (1 xter)
Industry	f1Ind.	f2Ind.	f1Ind.	f1Ind.	Industry Code (1 Digit); Industry Code (2 Digit); Industry Code (1 Digit); Industry Code (1 Digit)
Jobless	-	-	-	fJob.	Joblessness
LabSt	f96LFS.	f86LFS.	f91LFS.	f96LFS.	Labour Force Status

Variable Name	Format 1981	Format 1986	Format 1991	Format 1996	Variable Label
Link	flink.	flink.	flink.	flink.	Matched
MaoriAnc	-	-	f91Maor.	f96Maor.	Maori Ancestry Indicator
MaoriDes	f81Maor.	f86Maor.	-	-	Maori Descent Indicator
MarSt	f81Marr.	f86Marr.	f86Marr.	-	Marital Status
MarSt_L	-	-	-	f96MarL.	Marital Status (Legal)
MarSt_S	-	-	-	f96MarS.	Marital Status (Social)
NZDep91	-	-	fdeps.	-	NZ Deprivation 1991 scale
NZDep91sc	-	-	ftdep.	-	NZ Deprivation 1991 score (rounded)
NZDep96	fdeps.	fdeps.	-	fdeps.	NZ Deprivation 1996 scale
NZDep96sc	ftdep.	ftdep.	-	ftdep.	NZ Deprivation 1996 score (rounded)
NZDepFour	fdep4g.	fdep4g.	fdep4g.	fdep4g.	NZ Deprivation 1996 scale (4 groups); NZ Deprivation 1996 scale (4 groups); NZ Deprivation 1991 scale (4 groups); NZ Deprivation 1996 scale (4 groups)
O_EGP	fEGP.	fEGP.	fEGP.	fEGP.	EGP
O_EGPSp	-	fEGP.	-	-	EGP (Spouse)
O_ElleyIrv	fEI.	fEI.	fEI.	fEI.	Elley-Irving Index
O_ElleyIrvSp	-	fEI.	-	-	Elley-Irving Index (Spouse)
O_FarmFlg	fFarmF.	fFarmF.	fFarmF.	fFarmF.	Farmers Occupation Flag
O_FarmFlgFa	-	-	fFarmF.	-	Farmers Occupation Flag (Father)
O_FarmFlgMo	-	-	fFarmF.	-	Farmers Occupation Flag (Mother)
O_FarmFlgPr	-	-	fFarmF.	-	Farmers Occupation Flag (Parent)
O_FarmFlgSp	-	fFarmF.	-	-	Farmers Occupation Flag (Spouse)
O_Occ2X	f2xOcc.	f2xOcc.	f2xOcc.	f2xOcc.	Occupation Code - 2 Digits Occ68
O_OccSp2X	-	f2xOcc.	-	-	Spouse Occupation Code - 2 Digits Occ68
O_SEI91v	-	-	f91sei.	f91sei.	SEI 91 Values
O_SEI91vFa	-	-	f91sei.	-	SEI 91 Values (Father)
O_SEI91vMo	-	-	f91sei.	-	SEI 91 Values (Mother)
O_SEI91vPr	-	-	f91sei.	-	SEI 91 Values (Parent)
O_SEI96v	-	-	-	f96sei.	SEI 96 Values
PerType	fPRecT.	fPRecT.	fPRecT.	fPRecT.	Personal Record Type
PostAUIn	-	-	fPostC.	fPostC.	Post Census Area Unit Indicator
PreAUIn	-	-	fPreC.	fPreC.	Pre Census Area Unit Indicator
Religion	f81relg.	f81relg.	-	f81relg.	Religion - Main Groups; Religion - Treat Groups With Caution; Religion - Treat Groups With Caution
SeasDth	fseason.	fseason.	fseason.	fseason.	Season at Death

Variable Name	Format 1981	Format 1986	Format 1991	Format 1996	Variable Label
Sex	fvsex.	fvsex.	fvsex.	fvsex.	Sex
SexOc	fvsex.	fvsex.	-	-	Sex of Head of H/H; Sex of Occupier of H/H
SexPr	-	-	fvsex.	-	Sex of Parent
SmkCur	fSmkC.	-	-	-	Current Smoking Status
SmkEver	-	-	-	fSmkE.	Ever Smoked
SmkQnt	fSmkQ.	-	-	-	Quantity of Cigarettes Smoked in a day (23/3/81)
SmkReg	-	-	-	fSmkR.	Smoking Regularly
SmkStat	-	-	-	fSmkS.	Smoking Status
SocCap01	-	-	-	f01Soc.	Social Capital Index (0.1 steps)
SocCap40	-	-	-	f40Soc.	Social Capital Index (40 groups)
UsInd	f81USI.	fUSI.	fUSI.	fUSI.	Usual Residence Indicator
UsInd91	fUSI.	-	-	-	Usual Residence Indicator 1991
W_AgDepAdj	[8.6]	[8.6]	[8.6]	[8.6]	Deprivation Scaled Weight
W_AgEthAdj	[8.6]	[8.6]	[8.6]	[8.6]	Ethnicity Scaled Weight
W_AgICDAdj	[8.6]	[8.6]	[8.6]	[8.6]	Cause of Death Scaled Weight
W_Base	[8.6]	[8.6]	[8.6]	[8.6]	Base Linkage Weight
WgtStrata	[$21]	[$21]	[$21]	[$21]	Weight Stratum
YrsUR	-	f86YUR.	f91YUR.	f96YUR.	Years at Usual Residence

5.2.1 Person time

In order to protect the confidentiality of cohort members, date of death and date of birth could not be included on the cohort dataset. Person time of follow-up was therefore calculated based on the age in months at death and the age in months at census.

For all unlinked cohort members it was assumed that the individual was alive and living in New Zealand for the three years subsequent to the census. For these people the total person-time of follow-up was 36 months. For all linked cohort members the total person-time of follow-up was calculated as the age at death minus the age at census.

Most cohort analyses require controlling for age or age standardisation. In general, this control was done for five-year age groupings. As a consequence it was necessary to split the person-time of follow-up by age group. Four variables were created to facilitate analyses using Poisson regression. Two age categories were created corresponding to age at census, and age at the end of follow-up (age at death for linked cohort members and age three years after census for unlinked cohort members). The person-time was then split between the two age categories. Where the age categories at the census and at the end of follow-up were the same, the second category contained the value of missing. Any cohort member could contribute person-time to one or two age groups, as follow-up was for a maximum of three years and age was grouped into five-year bands. Because the age in months variable is the integer of the exact age in months it was assumed that the actual age for any value in months was the age in months plus 0.5. For example the actual age of a person for whom age in months equal to 712 months was assumed to be 712.5 months. Table 18 gives examples of the calculation of person-time by age categories.

Table 18: Examples of person-time calculations by age group.

Record number	*Age at census (months)* agec_cen_	*Age at end of follow up (months)*	*Age Category A (years)* AGECATA	*Age Category B (years)* AGETIMEA	*Person – time A (person–months)* AGECATB	*Person – time B (person-months)* AGETIMEB	*Linked* Linked
1	49	85	0 –4	5-9	10.5	25.5	0
2	306	342	15-19	.	36.0	.	0
3	816	841	65-69	70-74	23.5	1.5	1
4	121	127	10-14	.	6	.	1

To facilitate analyses where the age categories at census and end of follow-up were not the same the cohort record was split into two records for all analyses using Poisson regression.

Table 19 Illustration of splitting of records in Table 18 according to age category

Record number	*Age Category (years)*	*Person time (months)*	*Linked*
1	0-4	10.5	0
1	5-9	25.5	0
2	15-19	36	0
3	65-69	23.5	0
3	70-74	1.5	1
4	10-14	6	1

5.2.2 Ethnicity

5.2.2.1 Census

Census ethnicity was specified as either sole or prioritised ethnic group. Both classification systems were used for unlock and cohort analyses. Census ethnicity was not relevant to the bias analyses, as these analyses used mortality data only.

5.2.2.2 Mortality

For the weighting of census records (for linkage bias), we used the single option NMDS ethnicity variable for the first three census-mortality cohorts and the prioritised version of the NMDS ethnicity for 1996-99. (In the original 1991 bias analyses we used the NHI file ethnicity.(Blakely et al. 1999)

For the unlock analyses, a range of the mortality data ethnicity categorisations were compared with census categorisations (see future report).

5.2.3 Income

5.2.3.1 Equivalised income

'Equivalised total household income' was used as the measure of income for each individual in the NZCMS. This equivalised income was calculated from the *total household income* (i.e. the sum of personal incomes for individual household members) *'equivalised' for the number of household members*. This process is more complex than simply calculating the average income per household member, as it takes account of household economies of scale and the ratio of adults and children within each household. The revised Jensen Index was used to calculate equivalised incomes as well as simply dividing the total household income by the square root of the number of people in the household – this process is explained in more detail elsewhere.(Blakely 2002)(Blakely 2001)

5.2.3.2 Personal Income

For each of the four censuses, respondents were asked to identify their personal income as one of several categories, where each category referred to an income range (e.g. $10,001 to $15,000 per year). The *midpoint income* within this category was then used

as an estimate of that respondent's personal income (e.g. for the category above this would be $12,500 per year). Total household income was calculated by summing the midpoint income for each individual in that household.

5.2.4 Small area deprivation

The New Zealand small area deprivation index (NZDep) provides a population-based measure of socio-economic deprivation. The index uses a scale of 1 (least deprived) to 10 (most deprived). This is based on 10 variables (collected on census forms) that reflect material or social deprivation, including income, access to transport, living space, home ownership, employment, qualifications and support. The development of the NZDep is described elsewhere.(Crampton et al. 1997)

Individuals are assigned an NZDep score according to their address of usual residence. Records from the 1991 census cohort were assigned a score using the NZDep91 index (based on 1991 census); for all other cohorts, the NZDep96 index (based on 1996 census) was used. Cohort analyses were based on decile and quintile categories of NZDep.

5.2.5 Educational qualification

5.2.5.1 Highest School

Highest school qualification refers to the highest qualification gained while at school. Changes in school qualifications over time mean that older generations have different qualifications to younger generations.

5.2.5.2 Highest Qualification

This variable is the highest qualification gained including both school and post school qualifications. The value is derived by SNZ from the school and tertiary level qualification variables.

5.2.6 Labour force status

Labour force status is a measure of an individual's relationship to the formal work force. A person is in the labour force if they are either currently employed or if they are unemployed and actively seeking and available for work. Individuals who are not currently employed and not seeking work are not in the labour force. At the highest level of aggregation this variable has two possible values – in the labour force and not in the labour force.

5.2.7 Occupation Codes (NZSCO)

Occupation is recorded on census records using the New Zealand Standard Classification of Occupations (NZSCO). There are three versions of this classification system that applied during the 1980s and 1990s: one designed in 1968 (NZSCO68) that was coded for all four censuses, one developed in 1990 (NZSCO90) that was coded on the 1991 and 1996 census, and one developed in 1995 (NZSCO95) that was coded on the 1996 census. Thus the NZSCO68 code applied to all four census cohorts, while the others are used only in those censuses that were conducted after their development.

Occupation on mortality data was recorded using the same NZSCO coding system. The transition to the NZSCO90 system occurred during 1991, rendering all mortality data for 1991 unusable.(Blakely 2002)(Blakely 2001)

5.2.8 Occupational class

Occupational class is a socio-economic variable derived from standard classification of occupation codes. Three occupational class classifications are used in the NZCMS.

1. The Elley Irving occupational class is derived from the three digit NZSCO68.
2. When the NZSCO68 was superseded by the NZSCO90 classification of occupations the New Zealand socio-economic index (NZSEI) was developed to rank occupations according to education and income. The NZSEI relates to the NZSCO90 score and scale rank occupational codes.
3. The EGP classification scheme is an international classification based on the 1968 International classification of occupations from which the NZSCO68 was derived. This scheme is derived from the four digit occupational codes.

The New Zealand occupational class schemes do not differentiate farmers from other occupational classes. In order to facilitate comparisons with international data that differentiate farmers from other occupations a separate classification was developed for the NZSEI that separates farmers into a separate occupational class category (NZSCO90 codes 611 and 612). There is no direct concordance between these codes and three digit NZSCO68 or ISCO68 codes. However a direct one to many concordance does exist at the four digit level. A farmers flag was developed that identified these four digit codes to enable analyses that separated farmers into a separate occupational class.

5.2.9 Region

5.2.9.1 Meshblocks and Area Units

Meshblocks and area unit codes will not be used in cohort analyses per se. However, they will be retained on the SNZ master file to allow re-categorisation of the data by regions as required in the future.

5.2.9.2 Regions

Local government areas (n = 76) provide a useful number of regions for regional comparisons. For example, they can be used in income inequality analyses. Other options in Regional Councils (n=14) and regional Health authorities (n=4; used in unlock analyses extensively).

Should detailed analyses of regional variation in mortality be carried out in future, there will need to be an accompanying analysis of linkage bias by the equivalent regions.

5.2.9.3 Rurality

Census records were assigned to large urban, minor urban and rural codes. Due to poorer quality geocodes for rural addresses, any future analyses that compared mortality rates by rurality will need to adjust for residual linkage bias.

5.2.10 Death

5.2.10.1 Cause of death

In addition to analysing cohorts according to *all-cause mortality,* analyses were also conducted using *cause-specific mortality* as the outcome measure. The cause of death is recorded on each mortality record according to the International Cause of Death classification (Version 9). Cause-specific deaths were grouped according to the following table:

Table 20: ICD codes for grouping of cause-specific deaths

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

5.2.10.2 Season at death

Season at death was assigned according to the month of death, according to the following convention:

Spring	1 September to 30 November
Summer	1 December to 28 (or 29) February
Autumn	1 March to 31 May
Winter	1 June to 31 August

REFERENCES

Ajwani, S., Blakely, T., Robson, B., Atkinson, J., Fawcett, J., & Kiro, C. (2002). Estimating the numerator-denominator bias for the 1980s and 1990s. NZCMS Technical Report No. 4. (Also at http://www.wnmeds.ac.nz/nzcms-info.html). Wellington: Department of Public Health, Wellington School of Medicine and Health Sciences, University of Otago.

Blakely, T. (2001). *Socio-economic factors and mortality among 25-64 year olds: The New Zealand Census-Mortality Study. (Also at* http://www.wnmeds.ac.nz/nzcms-info.html). Unpublished Doctorate, University of Otago.

Blakely, T., & Atkinson, J. (2001). Unlocking the Numerator-Denominator Bias, 1991-94 Deaths. NZCMS Technical Report No. 2, *(Also at* http://www.wnmeds.ac.nz/nzcms-info.html). Wellington: Department of Public Health, Wellington School of Medicine, University of Otago.

Blakely, T., Kiro, C., & Woodward, A. (2002a). Unlocking the numerator-denominator bias. II: Adjustments to mortality rates by ethnicity and deprivation during 1991-94. *nzmj*, 115, 43-48.

Blakely, T., Robson, B., Atkinson, J., Sporle, A., & Kiro, C. (2002b). Unlocking the numerator-denominator bias. I: Adjustment ratios by ethnicity for 1991-94 mortality data. *nzmj*, 115, 39-43.

Blakely, T., & Salmond, C. (in press). A method to calculate the positive predictive value of record linkage. *International Journal of Epidemiology*.

Blakely, T., Salmond, C., & Woodward, A. (1999). Anonymous record linkage of 1991 census records and 1991-94 mortality records: The New Zealand Census-Mortality Study (Also at http://www.wnmeds.ac.nz/nzcms-info.html). Wellington: Department of Public Health, Wellington School of Medicine, University of Otago.

Blakely, T., Salmond, C., & Woodward, A. (2000). Anonymous linkage of New Zealand mortality and Census data. *anzjph*, 24, 92-95.

Blakely, T., Woodward, A., Pearce, N., Salmond, C., Kiro, C., & Davis, P. (2002c). Socio-economic factors and mortality among 25-64 year olds followed from 1991 to 1994: The New Zealand Census-Mortality Study. *nzmj*, 115, 93-97.

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

Fawcett, J., Atkinson, J., & Blakely, T. (2002). Weighting the 81, 86, 91 & 96 census-mortality cohorts to adjust for linkage bias. NZCMS Technical Report No. 5. (Also at http://www.wnmeds.ac.nz/nzcms-info.html). Wellington: Department of Public Health, Wellington School of Medicine and Health Sciences, University of Otago.

Jaro, M. (1995). Probabilistic linkage of large public health data files. *sm*, 14, 491-498.

Lewis, C. (2002). Information Analyst, New Zealand Health Information Service.

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

Blakely, T. (2002). The New Zealand Census-Mortality Study: Socioeconomic inequalities and adult mortality 1991-94., *Also at http://www.wnmeds.ac.nz/nzcms-info.html* (pp. 258). Wellington: Ministry of Health.

APPENDIX

5.3. SAS formats for variables included in cohort file

AGE FORMATS

Variable : AgeC_mths — **Age at Census (months)** — **1981,1986,1991,1996**

Variable : AgeD_mths — **Age at Death (months)** — **1981,1986,1991,1996**

Note: Values in data-set are single months, not grouped

Format : f5AgM

```
 0-< 60=' 0- 4 yrs'     60-<120=' 5- 9 yrs'    120-<180='10-14 yrs'    180-<240='15-19 yrs'
240-<300='20-24 yrs'    300-<360='25-29 yrs'    360-<420='30-34 yrs'    420-<480='35-39 yrs'
480-<540='40-44 yrs'    540-<600='45-49 yrs'    600-<660='50-54 yrs'    660-<720='55-59 yrs'
720-<780='60-64 yrs'    780-<840='65-69 yrs'    840-<900='70-74 yrs'    900-<936='75-77 yrs'
936-<960='78-79 yrs'       .,999='Miss Age'
```

Variable : AgeC_yrs — **Age at Census (years)** — **1981,1986,1991,1996**

Note: Values in data-set are single years, not grouped

Format : f5AgY

```
 0 -  4=' 0- 4 yrs'      5 -  9=' 5- 9 yrs'     10 - 14='10-14 yrs'      15 - 19='15-19 yrs'
20 - 24='20-24 yrs'     25 - 29='25-29 yrs'     30 - 34='30-34 yrs'      35 - 39='35-39 yrs'
40 - 44='40-44 yrs'     45 - 49='45-49 yrs'     50 - 54='50-54 yrs'      55 - 59='55-59 yrs'
60 - 64='60-64 yrs'     65 - 69='65-69 yrs'     70 - 74='70-74 yrs'      75 - 77='75-77 yrs'
78 - 79='78-79 yrs'        .,99='Miss Age'
```

Variable : AgeC_5yr — **Age at Census (5 year age groups)** — **1981,1986,1991,1996**

Format : f5AgG

```
  0=' 0- 4 yrs'     5=' 5- 9 yrs'     10='10-14 yrs'     15='15-19 yrs'     20='20-24 yrs'
 25='25-29 yrs'    30='30-34 yrs'     35='35-39 yrs'     40='40-44 yrs'     45='45-49 yrs'
 50='50-54 yrs'    55='55-59 yrs'     60='60-64 yrs'     65='65-69 yrs'     70='70-74 yrs'
 75='75-77 yrs'    78='78-79 yrs'     99='Miss Age'
```

Variable : AgeC_Gp — **Age at Census (5 Std Groups)** — **1981,1986,1991,1996**

Format : fagec

```
  1=' 0-14 yrs'     2='15-24 yrs'     3='25-44 yrs'     4='45-64 yrs'     5='65-74 yrs'
  6='75-79 yrs'     .='Missing'
```

Variable : (not used at present) but could be by AgeD_Gp — **1981,1986,1991,1996**

Format : faged

```
  1=' 0-14 yrs'     2='15-24 yrs'     3='25-44 yrs'     4='45-64 yrs'     5='65-78 yrs'
  .='Missing'
```

SEX FORMAT

Variable : Sex	**Sex**	**1981,1986,1991,1996**
Variable : SexOc	**Sex of Head of Household**	**1981**
Variable : SexOc	**Sex of Occupier of Household**	**1986**
Variable : SexPr	**Sex of Parent**	**1991**

Format : fvsex

```
1='Males'        2='Females'
```

ETHNICITY FORMATS

Variable : EthCenGp3	**Ethnicity (3 Groups)**	**1981**
Variable : EthCenGp4	**Ethnicity (4 Groups)**	**1981**
Variable : EthCenPr3	**Ethnicity -Prioritised (3 Groups)**	**1986,1991,1996**
Variable : EthCenPr4	**Ethnicity -Prioritised (4 Groups)**	**1986,1991,1996**
Variable : EthCenSol3	**Ethnicity -Sole (3 Groups)**	**1986,1991,1996**
Variable : EthCenSol4	**Ethnicity -Sole (4 Groups)**	**1986,1991,1996**

Format : f4eth

```
1='Maori'                2='Pacific People'          3='NonMaoriNonPac'      4='Asian'
5='NonMaoriNonPacNonAs'  9='Missing'
```

Variable : EthCenGp6_A	**Ethnicity -A**	**1986,1991,1996**
Variable : EthCenGp6_B	**Ethnicity -B**	**1986,1991,1996**
Variable : EthCenGp6_C	**Ethnicity -C**	**1986,1991,1996**

Format : fdeth

```
1='NZ Maori'             2='Pacific People'          4='Asian'
6='NZ European/Pakeha'   7='All Other Groups'        9='Missing'
```

Variable : EthCenPr5	**Ethnicity -Prioritised (SNZ Grouping)**	**1991**
Variable : EthCenSol5	**Ethnicity -Sole (SNZ Grouping)**	**1991**

Format : fnhiraw

```
1='Maori'          2='Pacific People'      3='Asian'          4='Other'          5='European'
```

Variable : EthCenPr5	**Ethnicity -Prioritised (SNZ Grouping)**	**1996**

Format : fraw

```
1='European'       2='Maori'            3='Pacific People'       4='Asian'       5='Other'
```

Variable : EthCenDet	**Ethnicity -Detailed**	**1991**

Format : f91EthD

```
1='NZ European Only'                          2='NZ European/Other Europeans'
3='Other Europeans Only'                      4='European/NZ Maori'
5='European/Samoan'                           6='European/Cook Island Maori'
7='European/Tongan'                           8='European/Niuean'
9='European/Tokelauan'                       10='European/Other P.I. Polynesian'
11='European/Chinese'                        12='European/Indian'
13='European/Fijian'                         14='European/Other Single Ethnic Group'
15='NZ Maori Only'                           16='NZ Maori/Samoan'
17='NZ Maori/Cook Island Maori'              18='NZ Maori/Tongan'
19='NZ Maori/Niuean'                         20='NZ Maori/Tokelauan'
```

```
21='NZ Maori/Other P.I. Polynesian'                    22='NZ Maori/Chinese'
23='NZ Maori/Indian'                                   24='NZ Maori/Fijian'
25='NZ Maori/Other Single Ethnic Group'                26='Samoan Only'
27='Samoan/Cook Island Maori'                          28='Samoan/Tongan'
29='Samoan/Niuean'                                     30='Samoan/Tokelauan'
31='Samoan/Other P.I. Polynesian'                      32='Samoan/Chinese'
33='Samoan/Indian'                                     34='Samoan/Fijian'
35='Cook Island Maori Only'                            36='Cook Island Maori/Tongan'
37='Cook Island Maori/Niuean'                          38='Cook Island Maori/Tokelauan'
39='Cook Island Maori/Other P.I. Polynesians'          40='Cook Island Maori/Chinese'
41='Cook Island Maori/Indian'                          42='Cook Island Maori/Fijian'
43='Tongan Only'                                       44='Tongan/Niuean'
45='Tongan/Tokelauan'                                  46='Tongan/Other P.I. Polynesian'
47='Tongan/Chinese'                                    48='Tongan/Indian'
49='Tongan/Fijian'                                     50='Niuean Only'
51='Niuean/Tokelauan'                                  52='Niuean/Other P.I. Polynesian'
53='Niuean/Chinese'                                    54='Niuean/Indian'
55='Niuean/Fijian'                                     56='Tokelauan Only'
57='Tokelauan/Other P.I. Polynesian'                   58='Tokelauan/Chinese'
59='Tokelauan/Indian'                                  60='Tokelauan/Fijian'
61='Other Single P.I. Polynesians'                     62='Fijian Only'
63='Other Single Pacific Islanders (excludes Polynesians)'
64='Other Two Ethnic Groups (at least one is Pacific Islander)'
65='Chinese Only'                                      66='Indian Only'
67='Chinese/Indian'                                    68='Vietnamese Only'
69='Japanese Only'                                     70='Kampuchean Only'
71='Sri Lankan Only'                                   72='Other Single Ethnic Groups Only'
73='Other Combinations of Two Ethnic Groups'
74='Three Ethnic Groups (NZ Maori/Pacific Islander/Other)'
75='Three Ethnic Groups (NZ Maori/Not Pacific Islander/Other)'
76='Three Ethnic Groups (Pacific Islander/Not NZ Maori/Other)'
77='Three Ethnic Groups (Not NZ Maori/Not Pacific Islander/Other)'
99='Not Specified'                                      .='Not Applicable'
```

COUNTRY OF BIRTH FORMAT

Variable : BirthGp | **Country of Birth** | **1981,1986,1991,1996**

Format : fbthgp

```
1='Born NZ'        2='Born Australia'     3='Born British Isles'     5='Born Pacific Islands'
8='Born Asia'      9='Born Elsewhere'     .='Missing'
```

MAORI ANCESTRY OF DESCENT FORMATS

Variable : MaoriDes | **Maori Descent Indicator** | **1981**

Format : f81Maor

```
0='Non-Maori Descent'              1='Maori Descent'              .='Missing'
```

Variable : MaoriDes | **Maori Descent Indicator** | **1986**

Format : f86Maor

```
0='Non-Maori Origin'               1='Maori Origin'               .='Missing'
```

Variable : MaoriAnc | **Maori Ancestry Indicator** | **1991**

Format : f91Maor

```
1='No Maori Ancestry'              2='Don''t Know'                3='Maori Ancestry'
9='Not Specified'                  .='Not Applicable'
```

Variable : MaoriAnc | **Maori Ancestry Indicator** | **1996**

Format : f96Maor

```
1='Maori Ancestry'          2='No Maori Ancestry'          3='Don''t Know'
9='Not Specified'          .,8='Missing or Not Applicable'
```

EDUCATION FORMATS

Variable : EdLAllCur — **Current Education Attendance Level** — **1981**
Variable : EdLAllPst — **Past Education Attendance Level** — **1981**

Format : fAtLev

```
  0='No attendance at any places of tertiary education'
  1='Still attending primary/secondary school'
  2='University'
  3='Teachers College'
  4='Polytechnic/Technical Inst./Community College'
  5='Other'
  6='University plus Teachers College'
  7='University plus Polytechnic/Tech Inst./Com. Coll.'
  8='Other Combinations'
 .,9='Not Specified'
```

Variable : EdLAllHgh — **Highest Level of Education Attendance** — **1981**

Format : fEdAtt

```
  1='Still Attending'                  2='No Secondary'
  3='Secondary to 5th Form'            4='6th Form'
  5='7th Form'                         6='University'
  7='Teachers Training College'        8='Polytech/Tech Inst./Com. Coll.'
  9='University & Teachers College'   10='Univ./Polytech/Tech/Com. Coll.'
 11='Other Tertiary'                 .,99='Not Specified'
```

Variable : EdLSchHgh — **School Attendance Level** — **1981**

Format : fscat

```
1='No primary or secondary schooling'   2='Primary or Intermediate, Form 2 (Std 6) or below'
3='Form 3'                              4='Form 4'
5='Form 5'                              6='Form 6'
7='Form 7'                             .,9='Not Specified'
```

Variable : EdQAll_A — **First Grouped Qualification Gained** — **1981**
Variable : EdQAll_B — **Second Grouped Qualification Gained** — **1981**
Variable : EdQAll_C — **Third Grouped Qualification Gained** — **1981**
Variable : EdQAll_D — **Fourth Grouped Qualification Gained** — **1981**

Format : f81qual

```
  1='Still at School'                              2='Doctorate & Masterate'
  3='Bachelorate'                                  4='Post-Graduate Diplomas'
  5='Under-Graduate Diplomas & Certificates'       6='Non-University Qualifications'
  9='Unidentified or Not Specified'                .='Missing'
```

Variable : EdQAllHgh — **Highest Qualification Obtained** — **1981**

Format : f81Hqal

```
  1='Post Graduate Degree or Degree'
  2='Undergraduate Degree, NZ Cert/Diploma Both NZC & Techn. C,Techn. Cert,Teaching/Nursing'
  3='Trade Certificates, other Tertiary Qualification'
  4='Higher School Certificate/Bursary, Sixth Form Certificate'
  5='School Certificate'
  6='Other School Qualification'
  7='Still at School'
  8='No Qualification'
  9='Not Specified'
```

Variable : EdQAllHgh — **Highest Qualification Gained (SNZ Protocol)** — **1986**

Format : f86Hqal

```
  1='Postgraduate Degree or Degree'
  2='Undergraduate Degree, NZ Cert/Diploma Both NZC & Technical,Techn. Cert,Teacher/Nursing'
  3='Trade Certificates, other Tertiary Qualification'
  4='Higher School Certificate/Bursary,Sixth Form Certificate'
```

```
    5='School Certificate'
    6='Other School Qualification'
    7='Still at School'
    8='No Qualification'
    9='Not Specified'
```

Variable : EdQAllHgh **Highest Qualification Obtained** **1991**

Format : f91HQ

```
   1='Postgraduates Degree'
   2='Bachelars Degree'
   3='Under Graduate Certificate/Diploma'
   4='Technicians Certificate'
   5='Teachers/Nurses Certificate/Diploma'
   6='Trade Certificate'
   7='Other Tertiary Qualifications'
   8='University Bursary/Scholarship/Higher School Leaving Cert'
   9='Sixth Form Certificate/University Entrance'
  10='School Certificate'
  11='Other School Qualifications (includes Overseas)'
  12='Still at School'
  13='No Qualifications'
 .,14='Not Specified'
```

Variable : EdQAllHgh **Derived Highest Qualification Gained** **1996**

Format : f96Hqal

```
   1='School Certificate'                    2='Sixth Form Certificate'
   3='Higher School Certificate'             4='Bachelor Degree'
   5='Higher Degree'                         6='Other School Qualification'
   7='No Qualification'                      8='Other Post-School Qualification'
   9='Not Specified'
```

Variable : EdQAllHghDet **Highest Qualification Gained** **1996**

Format : f96HQ

```
 9='Higher Degree'                            8='Bachelar Degree'
 7='Advanced Vocational Qualification'        6='Intermediate Vocational Qualification'
 5='Skilled Vocational Qualification'         4='Basic Vocational Qualification'
88='Post School Qualification, not applicable'  87='Post School Qualification, unidentifiable'
89='Post School Qualification, not specified'    3='Higher School Qualification'
 2='Sixth Form Qualification'                    1='School Certificate Qualification'
74='Overseas School Qualification'           78='School Qualification, not applicable'
76='School Qualification, not identifiable'  79='School Qualification, not specified'
77='No Qualification'                       .,99='Not Specified'
```

Variable : EdQSchHgh **Highest School Qualification** **1981**

Format : f81sql

```
   0='No School Qualification'
   1='University Scholarship, or A or B Bursary'
   2='Higher School Certificate or Higher Leaving Cert'
   3='University Entrance, Matriculation'
   4='Endorsed School Cert, or Sixth Form Cert in >=4 subj'
   5='Sixth Form Certificate in 1, 2 or 3 subjects'
   6='School Certificate, or >=3 subject passes in School Cert subj'
   7='Pass in 1 or 2 School Certificate subjects'
   8='Other (must be valid qualifications)'
 .,9='Not Specified'
```

Variable : EdQSchHgh **Highest School Qualification** **1986**

Format : f86sql

```
   1='No School Qualification'               2='School Certificate, 1 or 2 Passes'
   3='School Certificate, >=3 Passes'        4='6th Form Certificate, Endorsed School Cert'
```

```
    5='University Entrance, Matriculation'        6='Higher School Cert or Higher Leaving Cert'
    7='University Bursary or Scholarship'         8='Other'
  .,9='Not Specified'
```

Variable : EdQSchHgh **Highest School Qualification** **1991**

Format : f91sql

```
    1='No School Qualifications'                    2='School Certificate (>=1 subjects)'
    3='Sixth Form Cert, Univ Entrance (>=1 subj)'   4='Higher School Cert, Higher Leaving Cert'
    5='University Bursary or Scholarship'           6='Overseas Qualification'
    7='Other School Qualification'                 .,9='Not Specified'
```

Variable : EdQSchHgh **Highest School Qualification** **1996**

Format : f96sql

```
   10='NZ School Certificate in >=1 subj'
   20='NZ Sixth Form Certificate in >=1 subj'
   30='NZ University Entrance before 1986 in >=1 subj'
   40='NZ Higher School Cert or Higher Leaving Cert'
   50='NZ University Bursary, Entrance or Scholarship'
   70='Overseas Secondary School Qual not further defined'
   71='Overseas Equivalent to School Certificate Qual'
   72='Overseas Equivalent to Sixth Form Qual'
   73='Overseas Equivalent to Higher School Qual'
   74='Other Overseas Qualification nec'
   88='Tertiary Qualification'
   98='Unidentifiable'
  .,99='Not Specified'
```

Variable : EdQTer_A **Tertiary Qual Gained, Group A** **1991**

Format : f91Tqa

```
    0='Neither Trade Cert/Advanced Trade Cert or Nursing Cert/Diploma'
    1='Trade Certificate/Advanced Trade Certificate'
    2='Both Trade Cert/Advanced Trade Cert and Nursing Cert/Diploma'
    3='Nursing Certificate'
    7='Still at School'
  .,9='Not Specified'
```

Variable : EdQTer_A **Tertiary Qual 1 Attainment Level** **1996**

Format : f96Ter

```
    0='Category of Attainment Not Stated'           1='School Certificate'
    2='Sixth Form Qualification'                    3='Higher School Qualification'
    4='Basic Vocational Qualification'              5='Skilled Vocational Qualification'
    6='Intermediate Vocational Qualification'       7='Advanced Vocational Qualification'
    8='Bachelors Degree'                            9='Higher Degree'
   88='Category of Attainment Unidentifiable'     .,99='Missing'
```

Variable : EdQTer_B **Tertiary Qual Gained, Group B** **1991**

Format : f91TQb

```
    0='Neither NZ certificate/Diploma or Technician Certificate'
    1='NZ Certificate/Diploma'
    2='Both NZ Certificate/Diploma and Technicians Certificate'
    3='Technicians Certificate'
    7='Still at School'
  .,9='Not Specified'
```

Variable : EdQTer_C **Tertiary Qual Gained, Group C** **1991**

Format : f91TQc

```
    0='Neither Teacher Cert/Diploma or University Cert/Diploma below Bachelor level'
    1='Teachers Certificate/Diploma'
    2='Both Teachers Certificate/Diploma and University'
    3='University certificate/Diploma below Bachelors Level'
    7='Still at School'
```

```
  .,9='Not Specified'
```

Variable : EdQTer_D **Tertiary Qual Gained, Group D** **1991**

Format : f91TQd

```
  0='Neither Bachelor Degree or Post Graduate Degree Cert/Diploma'
  1='Bachelors Degree'
  2='Both Bachelors Degree and Postgraduate Degree Cert/Diploma'
  3='Postgraduate Degree Certificate/Diploma'
  7='Still at School'
 .,9='Not Specified'
```

Variable : EdQTer_E **Tertiary Qual Gained, Group E** **1991**

Format : f91Tqe

```
0='No Other Qualifications'     1='???Unsure'     7='Still at School'      .,9='Not Specified'
```

Variable : EdQTerHgh **Tertiary Qualification Gained** **1981**

Format : f81TQ

```
  1='Still at School'                                2='No Qualification'
  3='Trade and Non-University Qualification'         4='Undergraduate'
  5='Bachelor and Postgraduate'                      6='Other'
  7='Not Specified'
```

Variable : EdQTerHgh **Tertiary Qualification Gained** **1986**

Format : f86tql

```
  2='Still at School, or No Qualifications'          3='Trade Certificate'
  4='Nursing Certificate/Diploma'                    5='Teachers Certificate/Diploma'
  6='Technicians Certificate'                        7='NZ Certificate/Diploma'
  8='Undergraduate Certificate/Diploma'              9='Baccalureate'
 10='Postgraduate Degree/Cert/Diploma'              11='Other'
 12='Two Tertiary Qualifications'                   13='Three or more Tertiary Qualifications'
.,99='Not Specified'
```

EMPLOYMENT FORMAT

Variable : EmpSt **Employment Status** **1981**

Format : f81Emp

```
 0='Self-Employed, employing labour'        1='Self-Employed, not employing labour'
 2='Wages or salary'                        3='Relative (unpaid) assisting in business'
 4='Unemployed & seeking work'              5='Not specified but working >=20 hours weekly'
 6='Retired'                                7='Full time student'
 8='Household duties (unpaid)'              9='Other persons not working for financial reward'
 .='Missing or Not Specified'
```

Variable : EmpSt **Employment Status** **1986**

Format : f86Emp

```
  1='Full-Time Labour Force:Self-Employed (Employees)'
  2='Full-Time Labour Force:Self-Employed (No Employees)'
  3='Full-Time Labour Force:Wage & Salary Earner'
  4='Full-Time Labour Force:Relative Assisting'
  5='Full-Time Labour Force:Unemployed'
  6='Full-Time Labour Force:Not Specified'
  7='Part-Time Labour Force:Self-Employed (Employees)'
  8='Part-Time Labour Force:Self-Employed (No Employees)'
  9='Part-Time Labour Force:Wage & Salary Earner'
 10='Part-Time Labour Force:Relative Assisting'
 11='Part-Time Labour Force:Unemployed'
 12='Part-Time Labour Force:Not Specified'
 13='Non Labour Force'
  .='Missing'
```

Variable : EmpSt **Employment Status** **1996**

Format : f96Emp

```
 1='Full-Time Wage & Salary Earner'               2='Full-Time Self-Employed (No Employees)'
 3='Full-Time Self-Employed (Employees)'          4='Full-Time Unpaid Family Worker'
 5='Full-Time Not Specified Status in Employment' 6='Part-Time Wage & Salary Earner'
 7='Part-Time Self-Employed (No Employees)'       8='Part-Time Self-Employed (Employees)'
 9='Part-Time Unpaid Family Worker'              10='Part-Time Not Specified Status in Employment'
11='Unemployed and Actively Seeking Work'        12='Not in Labour Force'
13='Labour Force Status Not Available'            .='Missing'
```

LABOUR FORCE STATUS FORMATS

Variable : Used on EmpSt to create LabSt **1986**

InFormat : I86LFS

```
1,2,3,4=1          7,8,9,10=2          5,11=3          13=4          6=7          12=8
```

Variable : LabSt **Labour Force Status** **1986**

Format : f86LFS

```
1='Employed Full-Time'          2='Employed Part-Time'              3='Unemployed'
4='Not in Labour Force'         7='Full-Time:Not Specified'         8='Part-Time:Not Specified'
.,9='Not Specified'
```

Variable : LabSt **Labour Force Status** **1981,1996**

Format : f96LFS

```
1='Employed Full-Time'            2='Employed Part-Time'              3='Unemployed'
4='Not in Labour Force'          .,9='Not Specified'
```

Variable : LabSt **Labour Force Status** **1991**

Format : f91LFS

```
1='Gainfully Employed in the Full-Time Labour Force'
2='Gainfully Employed in the Part-Time Labour Force'
3='Unemployed & Actively Seeking Full-Time Work'
4='Unemployed & Actively Seeking Part-Time Work'
5='Non Labour Force (Seeking Work but Not Available)'
6='Non Labour Force (Available for Work but Not Seeking)'
7='Non Labour Force (Not Seeking & Not Available)'
.='Not Applicable'
```

JOBLESSNESS FORMAT

Variable : Jobless **Joblessness** **1996**

Format : fJob

```
1='Jobless-Available & Actively Seeking Work'
2='Jobless-Available but Not Actively Seeking Work'
3='Jobless-Actively Seeking Work but Not Available'
4='Not Jobless-Working'
5='Not Jobless-Not Available & Not Actively Seeking'
9='Not Classifiable'
.,8='Missing'
```

HOURS WORKED FORMATS

Variable : HrsWk **Total Hours Worked (per week)** **1981**

Note: Values in data-set are single numbers, not grouped

Format : f81hwk

```
0='    0 hours per week'       1- 9=' 1- 9 hours per week'      10-19='10-19 hours per week'
20-29='20-29 hours per week'   30-39='30-39 hours per week'     40-49='40-49 hours per week'
50-59='50-59 hours per week'   60-69='60-69 hours per week'     70-79='70-79 hours per week'
80-89='80-89 hours per week'   90-96='90-96 hours per week'        97='>=97 hours per week'
.,98='Not Specified'
```

Variable : HrsWkG **Total Number of Hours Worked** **1996**

Format : fhwk

```
1='0 to <30 hours worked'        2='30 to <50 hours worked'        3='50 or more hours worked'
.='Missing Hours'
```

INCOME FORMATS

Variable : H_THInc **Total Household Income** **1981**
Variable : I_Tinc **Total Personal Income** **1981**

Format : f81Inc

```
  0='Nil Income'          1='     $1 -   $249'     2='   $250 -   $499'
  3='   $500 -   $999'    4=' $1,000 -  $1,999'    5=' $2,000 -  $3,499'
  6=' $3,500 -  $4,999'   7=' $5,000 -  $6,499'    8=' $6,500 -  $7,999'
  9=' $8,000 -  $9,999'  10='$10,000 - $11,999'   11='$12,000 - $13,999'
 12='$14,000 - $15,999'  13='$16,000 - $17,999'   14='$18,000 - $19,999'
 15='$20,000 - $22,499'  16='$22,500 - $24,999'   17='$25,000 - $27,499'
 18='$27,500 - $29,999'  19='$30,000 - $34,999'   20='$35,000 - $39,999'
 21='$40,000 - $49,999'  22='$50,000 - $59,999'   23='$60,000 and over'
97,98='Not Available'   .,99='Not Specified'
```

Variable : H_THInc **Total Household Income** **1986**
Variable : I_Tinc **Total Personal Income** **1986**

Format : f86Inc

```
  1='Nil or loss'         2='     $1 -  $1,000'    3=' $1,001 -  $2,500'
  4=' $2,501 -  $5,000'   5=' $5,001 -  $7,500'    6=' $7,501 - $10,000'
  7='$10,001 - $12,500'   8='$12,501 - $15,000'    9='$15,001 - $17,500'
 10='$17,501 - $20,000'  11='$20,001 - $25,000'   12='$25,001 - $30,000'
 13='$30,001 - $35,000'  14='$35,001 - $40,000'   15='$40,001 - $50,000'
 16='$50,001 and over'   98='Not Available'      .,99='Not Specified'
```

Variable : H_THInc **Total Household Income** **1991**
Variable : I_Tinc **Total Personal Income** **1991**

Format : f91Inc

```
  1='Nil or loss'         2='     $1 -  $2,500'    3=' $2,501 -  $5,000'
  4=' $5,001 -  $7,500'   5=' $7,501 - $10,000'    6='$10,001 - $15,000'
  7='$15,001 - $20,000'   8='$20,001 - $25,000'    9='$25,001 - $30,000'
 10='$30,001 - $40,000'  11='$40,001 - $50,000'   12='$50,001 - $70,000'
 13='$70,001 and over'   98='Not Available'      .,99='Not Specified'
```

Variable : H_THInc **Total Household Income** **1996**
Variable : I_Tinc **Total Personal Income** **1996**

Format : f96Inc

```
  1='Loss'                2='Zero Income'          3='     $1 -  $5,000'
  4=' $5,001 - $10,000'   5='$10,001 - $15,000'    6='$15,001 - $20,000'
  7='$20,001 - $25,000'   8='$25,001 - $30,000'    9='$30,001 - $40,000'
 10='$40,001 - $50,000'  11='$50,001 - $70,000'   12='$70,001 -$100,000'
 13='$100,001 and over'  88='Unidentifiable'      98='Not Available'
.,99='Not Specified'
```

SOURCE OF INCOME FORMATS

Variable : H_IncAC	H/H Inc. Srce - ACC Regular Payments	1996
Variable : H_IncDP	H/H Inc. Srce - Domestic Purposes Benefit	1996
Variable : H_IncGB	H/H Inc. Srce - Other Government Benefits	1996
Variable : H_IncIB	H/H Inc. Srce - Invalids Benefit	1996
Variable : H_IncSB	H/H Inc. Srce - Sickness Benefit	1996
Variable : H_IncSE	H/H Inc. Srce - Self-employment	1996
Variable : H_IncUB	H/H Inc. Srce - Unemployment Benefit	1996
Variable : H_IncWS	H/H Inc. Srce - Wages/Salary etc.	1996
Variable : I_PIS_AC	Personal Inc. Srce - ACC Regular Payments	1996
Variable : I_PIS_DB	Personal Inc. Srce - Domestic Purposes Benefit	1996
Variable : I_PIS_GB	Personal Inc. Srce - Other Government Benefit	1996
Variable : I_PIS_IB	Personal Inc. Srce - Invalids Benefit	1996
Variable : I_PIS_SB	Personal Inc. Srce - Sickness Benefit	1996
Variable : I_PIS_SE	Personal Inc. Srce - Self-employment	1996
Variable : I_PIS_UB	Personal Inc. Srce - Unemployment Benefit	1996
Variable : I_PIS_WS	Personal Inc. Srce - Wages/Salary etc.	1996

Format : fIncS

```
  1='Wages, salary, commissions, bonuses etc. paid by employer'
  2='Self-employment, or business you own and work in'
  3='Interest, dividends, rent, other investments'
  4='ACC regular payments'
  5='NZ superannuation'
  6='Other superannuation, pensions, annuities'
  7='Unemployment benefit'
  8='Domestic purposes benefit'
  9='Sickness benefit'
 10='Invalid''s benefit'
 11='Student allowance'
 12='Other govt benefits, income support payments or war pensions'
 13='Other sources of income'
.,99='None'
```

Variable : I_DPB	Domestic Purposes Benefit	1981,1986

Format : fiDPB

```
0='Did not receive DPB'      1='Received Domestic Purposes Benefit'      .='Not Applicable'
```

Variable : I_FamBen	Family Benefit	1981,1986

Format : fiFB

```
0='Did not receive FB'       1='Received Family Benefit'                 .='Not Applicable'
```

Variable : I_FamCare	Family Care Benefit	1986

Format : fiFC

```
0='Did not receive FC'       1='Received Family Care'                    .='Not Applicable'
```

Variable : I_IncSup	Income Support Payments Indicator	1981,1986

Format : fiIS

```
0='Did not receive any IS'   1='Received any Income Support'          .,9='Not Applicable'
```

Variable : I_InvalBen	Invalids Benefit Indicator	1981

Format : fiIB

```
0='Did not receive IB'       1='Received Invalids Benefit'               .='Not Applicable'
```

Variable : I_ISP_Der **Income Support Payments -Derived** **1991**

Format : fi91ISP

```
 1='Family Benefit'                                2='National Superannuation/GRI'
 3='Accident Compensation Weekly Payments'         4='Domestic Purposes Benefit'
 5='Unemployment Benefit'                          6='Youth and Student Allowance'
 7='Sickness/Invalids Benefit'                     8='War Pension'
 9='Other Support Payments'                       10='Family Benefit & Family Support'
11='Family Benefit & Domestic Purposes Benefit'   12='Other Combinations of >=2 payments'
13='No Payments Received'                         14='Not Specified'
 .='Not Applicable'
```

Variable : I_ISPA **Income Support Payment Group A** **1991**
Variable : I_ISPB **Income Support Payment Group B** **1991**
Variable : I_ISPC **Income Support Payment Group C** **1991**
Variable : I_ISPD **Income Support Payment Group D** **1991**
Variable : I_ISPE **Income Support Payment Group E** **1991**

Format : fi91IG

```
 0='Did Not Receive Payment for x'   1='Income Support Payment Code 1'
 2='Income Support Payment Code 2'   3='Income Support Payment Code 3'
 9='Not Specified'                   .='Not Applicable'
```

Variable : I_OispG **Other Income Support Payments -Grouped** **1991**

Format : fi91IO

```
11='No Other Income Support Payments'   12='Widows Pension Payments'
13='Disability Allowance Payments'      14='Maintenance from Former Partner'
10='All Other Income Support Payments'   .='Not Applicable'
```

Variable : I_SickBen **Sickness Benefit** **1981,1986**

Format : fiSick

```
 0='Did not receive SB'   1='Received Sickness Benefit'   .='Not Applicable'
```

Variable : I_UnEmpBen **Unemployment Benefit** **1981,1986**

Format : fiUB

```
 0='Did not receive UB'   1='Received Unemployment Benefit'   .='Not Applicable'
```

MARITAL STATUS FORMATS

Variable : MarSt | **Marital Status** | **1981**

Format : f81Marr

```
1='Never Married'          2='Married'             3='Separated'                4='Widowed'
5='Divorced'               9='Not Specified'       .='Missing or Not Applicable'
```

Variable : MarSt | **Marital Status** | **1986,1991**

Format : f86Marr

```
1='Never Married'            2='Married, First Time'          3='Remarried'
4='Separated'                5='Divorced'                    6='Widowed'
9='Not Specified'          .,8='Missing or Not Applicable'
```

Variable : MarSt_L | **Marital Status (Legal)** | **1996**

Format : f96MarL

```
111='Married (not separated)-First Marriage'
121='Married (not separated)-Subsequent Marriage'
131='Married (not separated)-Not Further Classifiable'
211='Never Married'
221='Separated'
222='Divorced'
223='Widowed'
.,911='Not Specified'
```

Variable : MarSt_S | **Marital Status (Social)** | **1996**

Format : f96MarS

```
111='Partnered, Legal Spouse (not separated)-First Marriage'
112='Partnered, Legal Spouse (not separated)-Subsequent Marriage'
113='Partnered, Legal Spouse (not separated)-Not Further Classifiable'
121='Partnered, De Facto Spouse-Opposite-sex Couple'
122='Partnered, De Facto Spouse-Same-sex Couple'
131='Partnered, Not Further Classifiable'
211='Non-partnered, Never Married'
221='Non-partnered, Separated'
222='Non-partnered, Divorced'
223='Non-partnered, Widowed'
.,911='Not Specified'
```

BABY BORN FORMAT

Variable : BabyBrn | **Number of Live Babies Given Birth To** | **1996**

Format : fBBrn

```
0='No Children'          1='1 Child'              2='2 Children'             3='3 Children'
4='4 Children'           5='5 Children'           6='6 Children'             7='7 Children'
8='8 Children'           9='9 Children'          10='10 or More Children'
88='Unidentifiable'     98='Object to Answering'  99='Not Specified'
```

FAMILY FORMATS

Variable : FamCode **Family Code** **1981,1986**

Format : fFamC

```
0='Parent (1st Family)'        1='Child (1st Family)'         2='Parent (2nd Family)'
3='Child (2nd Family)'         4='Parent (3rd Family)'        5='Child (3rd Family)'
6='Member (Other Families)'     7='Non Family Person'          8='Person Alone'
9='Guest or Visitor'         .,99='Not Applicable'
```

Variable : FamType **Family Type** **1991**

Format : f91FamT

```
 1='One Parent Family with Dependent Children Only'
 2='One Parent Family with Dependent & Adult Children'
 3='One Parent Family with Adult Children Only'
 4='Two Parent Family with Dependent Children Only (Youngest <= 0-4 Yrs)'
 5='Two Parent Family with Dependent Children Only (Youngest <= 5-12 Yrs'
 6='Two Parent Family with Dependent Children Only (Youngest <= 13-15 Yrs'
 7='Two Parent Family with Dependent Children Only (Youngest <= 16-18 Yrs)'
 8='Two Parent Family with Dependent & Adult Children (Youngest <= 0-4 Yrs)'
 9='Two Parent Family with Dependent & Adult Children (Youngest <= 5-12 Yrs)'
10='Two Parent Family with Dependent & Adult Children (Youngest <= 13-15 Yrs)'
11='Two Parent Family with Dependent & Adult Children (Youngest <= 16-18 Yrs)'
12='Two Parent Family with Adult Children Only'
13='Couple Only with Wife Aged 0-29 Years'
14='Couple Only with Wife Aged 30-44 Years'
15='Couple Only with Wife Aged 45-59 Years'
16='Couple Only with Wife Aged >=60 Years'
17='Non Family Unit'
18='Unknown Coding Value'
 .='Missing'
```

Variable : FamType **Family Type** **1996**

Format : f96FamT

```
11='Couple without children'
21='Couple with dependent children only'
22='Couple with dependent and adult children'
23='Couple with adult children only'
29='Couple with children, dependency status not classifiable'
31='One parent family with dependent children only'
32='One parent family with dependent and adult children'
33='One parent family with adult children only'
39='One parent family with children, dependency status not classifiable'
91='Family type not classifiable'
 .='Missing'
```

GEOGRAPHICAL VARIABLES

Variable : G_AHB — **Area Health Board 1989** — **1981,1986,1991,1996**

Format : f89AHB

```
  1='Northland'              2='Auckland'              3='Waikato'
  4='Bay of Plenty'          5='Tairawhiti'            6='Hawke''s Bay'
  7='Taranaki'               8='Manawatu/Wanganui'     9='Wellington'
 10='Nelson/Marlborough'    12='West Coast'           13='Canterbury'
 14='Otago'                 15='Southland'            88='Overseas'
 99='Not Applicable'         .='Missing'
```

Area Health Districts 1993 based on 1995 TLAs

Variable : G_AHD — **Area Health District 1993** — **1981,1986,1996**

Format : f93AHD

```
  1='Northland'              2='North West Auckland'   3='Central Auckland'
  4='South Auckland'         5='Eastern Bay of Plenty' 6='Rotorua'
  7='Taupo'                  8='Tauranga'              9='Gisborne'
 10='Taranaki'              11='Waiakto'              12='Ruapehu'
 13='Wanganui'              14='Manawatu'             15='Hawkes Bay'
 16='Wairarapa'             17='Hutt'                 18='Wellington'
 19='Nelson-Marlborough'    20='West Coast'           21='Canterbury'
 22='South Canterbury'      23='Otago'                24='Southland'
.,99='Not Applicable'
```

Variable : G_AHD — **Area Health District 1993** — **1991**

Variable : G_AHBD91 — **Usual Residence Area Health Board Consituent District** — **1991**

Format : f91AHD

```
 101='Maungataniwha'                          102='Bay of Islands'
 103='Kaipara'                                104='Whangarei Rural'
 105='Whangarei Urban'                        198='Northland, not further defined'
 201='Rodney/North Shore'                     202='Waitakere'
 203='Auckland'                               204='Manukau'
 205='Papakura/Franklin'                      298='Auckland, not further defined'
 301='Thames-Coromandel'                      302='North Waikato'
 303='Waihou'                                 304='Hamilton West'
 305='Hamilton East'                          306='Northern King Country'
 307='Waipa'                                  308='South Waikato'
 309='Taupo'                                  398='Waikato, not further defined'
 401='Western Bay of Plenty'                  402='Tauranga'
 403='Rotorua'                                404='Eastern Bay of Plenty'
 498='Bay of Plenty, not further defined'     501='Waiapu'
 502='Cook'                                   503='Gisborne'
598='Tairawhiti, not further defined'         601='Wairoa'
602='Ngaruroro'                               603='Napier'
604='Hastings'                                605='Central Hawke''s Bay'
698='Hawke''s Bay, not further defined'       701='North Taranaki'
702='New Plymouth'                            703='Stratford'
704='South Taranaki'                          798='Taranaki, not further defined'
 801='Wanganui'                               802='Rangitikei'
 803='Manawatu'                               804='Palmerston North'
 805='Tararua'                                806='Horowhenua'
 898='Manawatu/Wanganui, not further defined' 901='Kapiti Coast'
 902='Porirua'                                903='Upper Hutt'
 904='Lower Hutt'                             905='Wellington North'
 906='Wellington South'                       907='Wairarapa'
 998='Wellington, not further defined'
1001='Tasman'                                1002='Nelson'
1003='Marlborough'                           1098='Marlborough, not further defined'
```

```
1201='Buller'                                   1202='Grey'
1203='Westland'                                 1298='West Coast, not further defined'
1301='North Canterbury'                         1302='Fitzgerald'
1303='Godley'                                   1304='Ashburton'
1305='South Canterbury/Waitaki'                 1398='Canterbury, not further defined'
1401='Dunstan'                                  1402='Moeraki'
1403='Molyneux'                                 1404='Cargill'
1405='Wickcliffe'                               1498='Otago, not further defined'
1501='Te Anau'                                  1502='Hokonui'
1503='Gore'                                     1504='Waikiwi'
1505='Awarua'                                   1506='Dome'
1598='Southland, not further defined'
.,9696,9898,9999='Not Applicable or Not Specified'
```

Variable : G_RHA — **Regional Health Authority (1989 AHB)** — **1981,1986,1991,1996**

Format : frha

```
1='Northern'        2='Midland'        3='Central'        4='Southern'
```

Variable : G_Rurality — **Rurality Indicator** — **1981,1986,1991,1996**

Format : frural

```
1='Urban'          2='Minor Urban'       3='Rural & Other'
```

Variable : G_TLA5yr — **TLA 1995 Address 5 Years Ago** — **1996**
Variable : G_TLA89 — **Territorial Local Authority 1989** — **1991**
Variable : G_TLA95 — **Territorial Local Authority 1995** — **1981,1986,1996**

Format : f95tla

```
     1='Far North'                2='Whangarei'                3='Kaipara'
     4='Rodney'                   5='North Shore'              6='Waitakere'
     7='Auckland'                 8='Manukau'                  9='Papakura'
    10='Franklin'                11='Thames Coromandel'       12='Hauraki'
    13='Waikato'                 15='Matamata-Piako'          16='Hamilton'
    17='Waipa'                   18='Otorohanga'              19='South Waikato'
    20='Waitomo'                 21='Taupo'                   22='Western Bay of Plenty'
    23='Tauranga'                24='Rotorua'                 25='Whakatane'
    26='Kawerau'                 27='Opotiki'                 28='Gisborne'
    29='Wairoa'                  30='Hastings'                31='Napier'
    32='Central Hawkes Bay'      33='New Plymouth'            34='Stratford'
    35='South Taranaki'          36='Ruapehu'                 37='Wanganui'
    38='Rangitikei'              39='Manawatu'                40='Palmerston North'
    41='Tararua'                 42='Horowhenua'              43='Kapiti Coast'
    44='Porirua'                 45='Upper Hutt'              46='Lower Hutt'
    47='Wellington'              48='Masterton'               49='Carterton'
    50='South Wairarapa'         51='Tasman'                  52='Nelson'
    53='Marlborough'             54='Kaikoura'                55='Buller'
    56='Grey'                    57='Westland'                58='Hurunui'
    59='Waimakariri'             60='Christchurch'            61='Banks Peninsula'
    62='Selwyn'                  63='Ashburton'               64='Timaru'
    65='Mackenzie'               66='Waimate'                 67='Chatham Islands'
    68='Waitaki'                 69='Central Otago'           70='Queenstown-Lakes'
    71='Dunedin'                 72='Clutha'                  73='Southland'
    74='Gore'                    75='Invercargill'           888='Overseas'
901-998='Other Groupings (N/A)' 999='TLA Not Applicable'      .='Missing'
```

Variable : G_UA91 **Usual Residence Urban Area 1991** **1991**

Format : f91UA

```
   1='Whangarei'                2='Northern Auckland Zone'       3='Western Auckland Zone'
   4='Central Auckland Zone'    5='Southern Auckland Zone'       6='Hamilton Zone'
  46='Cambridge Zone'          47='Te Awamutu Zone'             7='Tauranga'
   8='Rotorua'                  9='Gisborne'                    10='Napier'
  11='Hastings'                12='New Plymouth'               13='Wanganui'
  14='Palmerston North'        15='Upper Hutt Zone'            16='Lower Hutt Zone'
  17='Porirua Basin Zone'      18='Wellington City Zone'       19='Nelson'
  20='Christchurch'            22='Dunedin'                    23='Invercargill'
  24='Pukekohe'                25='Tokoroa'                    26='Taupo'
  27='Whakatane'               28='Hawera'                     29='Feilding'
  30='Levin'                   31='Kapiti'                     32='Masterton'
  33='Blenheim'                34='Greymouth'                  35='Ashburton'
  21='Timaru'                  36='Oamaru'                     37='Gore'
  38='Minor Urban Areas'       39='Shipping'                 40,41='Rural Areas'
42-45='Oceanic/Inlet'          96='No Fixed Abode'         .,98,99='Not Specified N.Z.'
```

Variable : G_UA96 **Usual Residence Urban Area 1996** **1981,1986,1996**

Format : f96UA

```
   1='Whangarei'                2='Northern Auckland Zone'       3='Western Auckland Zone'
   4='Central Auckland Zone'    5='Southern Auckland Zone'       6='Hamilton Zone'
   7='Cambridge Zone'           8='Te Awamutu Zone'              9='Tauranga'
  10='Rotorua'                 11='Gisborne'                   12='Napier Zone'
  13='Hastings Zone'           14='New Plymouth'               15='Wanganui'
  16='Palmerston North'        17='Upper Hutt Zone'            18='Lower Hutt Zone'
  19='Porirua Zone'            20='Wellington Zone'            21='Nelson'
  22='Christchurch'            23='Dunedin'                    24='Invercargill'
 101='Pukekohe'               102='Tokoroa'                   103='Taupo'
 104='Whakatane'              105='Hawera'                    106='Feilding'
 107='Levin'                  108='Kapiti'                    109='Masterton'
 110='Blenheim'               111='Greymouth'                 112='Ashburton'
 113='Timaru'                 114='Oamaru'                    115='Gore'
 201='Taipa Bay-Mangonui'     202='Kaitaia'                   203='Kerikeri'
 204='Russell'                205='Paihia'                    206='Kawakawa'
 207='Moerewa'                208='Kaikohe'                   209='Dargaville'
 210='Wellsford'              211='Warkworth'                 212='Snells Beach'
 213='Helensville'            214='Waiheke Island'            215='Waiuku'
 216='Raglan'                 217='Huntly'                    218='Otorohanga'
 219='Te Kuiti'               220='Taumarunui'                221='Whitianga'
 222='Coromandel'             223='Whangamata'                224='Tairua'
 225='Pauanui Beach'          226='Thames'                    227='Waihi Beach'
 228='Paeroa'                 229='Waihi'                     230='Te Aroha'
 231='Morrinsville'           232='Matamata'                  233='Putaruru'
 234='Katikati Community'     235='Te Puke Community'         236='Mangakino'
 237='Turangi'                238='Edgecumbe Community'       239='Kawerau'
 240='Murupara'               241='Opotiki'                   242='Wairoa'
 243='Waipawa'                244='Waipukurau'                245='Dannevirke'
 246='Woodville'              247='Waitara'                   248='Inglewood'
 249='Stratford'              250='Opunake'                   251='Eltham'
 252='Manaia'                 253='Patea'                     254='Ohakune'
 255='Raetihi'                256='Waiouru'                   257='Bulls'
 258='Taihape'                259='Marton'                    260='Foxton Community'
 261='Shannon'                262='Otaki'                     263='Pahiatua'
 264='Carterton'              265='Greytown'                  266='Featherston'
 267='Martinborough'          268='Picton'                    269='Kaikoura'
 270='Takaka'                 271='Brightwater'               272='Wakefield'
 273='Motueka'                274='Westport'                  275='Reefton'
 276='Hokitika'               277='Hanmer Springs'            278='Woodend'
 279='Rangiora'               280='Oxford'                    281='Darfield'
 282='Lincoln'                283='Leeston'                   284='Pleasant Point'
 285='Geraldine'              286='Temuka'                    287='Twizel Community'
```

```
    288='Waimate'                    289='Milton'                    290='Balclutha'
    291='Alexandra'                  292='Cromwell'                  293='Wanaka'
    294='Arrowtown'                  295='Queenstown'                296='Winton'
    297='Bluff'                      298='Te Anau'                   299='Riverton'
    501='Rural Centre'                              502='Rural (Incl. Some Off-Shore Islands)'
    505='Inland Water Not in Urban Area'            506='Inlet-Not in TLA'
    507='Inlet-In TLA but Not in Urban Area'        510='Oceanic-In Region But Not in TLA'
    511='Oceanic-Outside Region'                    888='Overseas'
  .,999='Urban Area Not Applicable'
```

NEW ZEALAND DEPRIVATION FORMATS

Variable : NZDep91 — **NZ Deprivation 1991 scale** — **1991**
Variable : NZDep96 — **NZ Deprivation 1996 scale** — **1981,1986,1996**

Format : fdeps

```
  1='Dep 1'      2='Dep 2'      3='Dep 3'      4='Dep 4'       5='Dep 5'      6='Dep 6'
  7='Dep 7'      8='Dep 8'      9='Dep 9'     10='Dep10'      0,.='Miss Dep'
```

Variable : NZDep91sc — **NZ Deprivation 1991 score (rounded)** — **1991**
Variable : NZDep96sc — **NZ Deprivation 1996 score (rounded)** — **1981,1986,1996**

Note: Values in data-set are single values, not grouped

Format : ftdep

```
       0='0 dep'              830- 899=' 830- 899 dep'        900- 999=' 900- 999 dep'
1000-1099='1000-1099 dep'    1100-1199='1100-1199 dep'      1200-1299='1200-1299 dep'
1300-1399='1300-1399 dep'    1400-1499='1400-1499 dep'      1500-1519='1500-1519 dep'
     1520='1520 dep'              1530='1530 dep'                  .='Missing dep'
```

Variable : NZDepFour — **NZ Deprivation 1991 scale (4 groups)** — **1991**
Variable : NZDepFour — **NZ Deprivation 1996 scale (4 groups)** — **1981,1986,1991**

Format : fdep4g

```
  1='Dep 1-4'      2='Dep 5-6'      3='Dep 7-8'      4='Dep 9-10'      0,.='Miss Dep'
```

SOCIAL CAPITAL FORMATS

Variable : Used to create SocCap01 **1996**

InFormat : isocn

```
-7.05-<-2.05=-450       -2.05-<-1.65=-185       -1.65-<-1.45=-155       -1.45-<-1.35=-140
-1.35-<-1.25=-130       -1.25-<-1.15=-120       -1.15-<-1.05=-110       -1.05-<-0.95=-100
-0.95-<-0.85= -90       -0.85-<-0.75= -80       -0.75-<-0.65= -70       -0.65-<-0.55= -60
-0.55-<-0.45= -50       -0.45-<-0.35= -40       -0.35-<-0.25= -30       -0.25-<-0.15= -20
-0.15-<-0.05= -10       -0.05-< 0.05=   0        0.05-< 0.15=  10        0.15-< 0.25=  20
 0.25-< 0.35=  30        0.35-< 0.45=  40        0.45-< 0.55=  50        0.55-< 0.65=  60
 0.65-< 0.75=  70        0.75-< 0.85=  80        0.85-< 0.95=  90        0.95-< 1.05= 100
 1.05-< 1.15= 110        1.15-< 1.25= 120        1.25-< 1.35= 130        1.35-< 1.45= 140
 1.45-< 1.65= 155        1.65-< 2.05= 185        2.05-  7.05= 450
```

Variable : SocCap01 **Social Capital Index (0.1 steps)** **1996**

Format : f01Soc

```
-450='-7.05 -< -2.05'       -185='-2.05 -< -1.65'       -155='-1.65 -< -1.45'
-140='-1.45 -< -1.35'       -130='-1.35 -< -1.25'       -120='-1.25 -< -1.15'
-110='-1.15 -< -1.05'       -100='-1.05 -< -0.95'        -90='-0.95 -< -0.85'
 -80='-0.85 -< -0.75'        -70='-0.75 -< -0.65'        -60='-0.65 -< -0.55'
 -50='-0.55 -< -0.45'        -40='-0.45 -< -0.35'        -30='-0.35 -< -0.25'
 -20='-0.25 -< -0.15'        -10='-0.15 -< -0.05'          0='-0.05 -<  0.05'
  10=' 0.05 -<  0.15'         20=' 0.15 -<  0.25'         30=' 0.25 -<  0.35'
  40=' 0.35 -<  0.45'         50=' 0.45 -<  0.55'         60=' 0.55 -<  0.65'
  70=' 0.65 -<  0.75'         80=' 0.75 -<  0.85'         90=' 0.85 -<  0.95'
 100=' 0.95 -<  1.05'        110=' 1.05 -<  1.15'        120=' 1.15 -<  1.25'
 130=' 1.25 -<  1.35'        140=' 1.35 -<  1.45'        155=' 1.45 -<  1.65'
 185=' 1.65 -<  2.05'        450=' 2.05 -   7.05'
```

Social Capital in 40 almost equal groupings (using RANK)

Variable : Used to create SocCap40 **1996**

InFormat : isocc

```
-4.000  -< -1.715  = 1       -1.715  -< -1.450  = 2        -1.450  -< -1.3195 = 3
-1.3195 -< -1.200  = 4       -1.200  -< -1.08977= 5        -1.08977-< -0.973  = 6
-0.973  -< -0.908  = 7       -0.908  -< -0.8334 = 8        -0.8334 -< -0.770  = 9
-0.770  -< -0.697  =10       -0.697  -< -0.5965 =11        -0.5965 -< -0.526  =12
-0.526  -< -0.460  =13       -0.460  -< -0.40283=14        -0.40283-< -0.337  =15
-0.337  -< -0.283  =16       -0.283  -< -0.2156 =17        -0.2156 -< -0.1605 =18
-0.1605 -< -0.10838=19       -0.10838-< -0.0594 =20        -0.0594 -<  0.009  =21
 0.009  -<  0.050  =22        0.050  -<  0.1011 =23         0.1011 -<  0.162  =24
 0.162  -<  0.234  =25        0.234  -<  0.281  =26         0.281  -<  0.34585=27
 0.34585-<  0.4142 =28        0.4142 -<  0.5008 =29         0.5008 -<  0.599  =30
 0.599  -<  0.701  =31        0.701  -<  0.785  =32         0.785  -<  0.857  =33
 0.857  -<  0.9613 =34        0.9613 -<  1.109  =35         1.109  -<  1.250  =36
 1.250  -<  1.45   =37        1.45   -<  1.7459 =38         1.7459 -<  2.099  =39
 2.099  -<  6.9    =40                    other=.
```

Variable : SocCap40 **Social Capital Index (40 groups)** **1996**

Format : f40Soc

```
 1='Soc Cap Group  1'    2='Soc Cap Group  2'     3='Soc Cap Group  3'     4='Soc Cap Group  4'
 5='Soc Cap Group  5'    6='Soc Cap Group  6'     7='Soc Cap Group  7'     8='Soc Cap Group  8'
 9='Soc Cap Group  9'   10='Soc Cap Group 10'    11='Soc Cap Group 11'    12='Soc Cap Group 12'
13='Soc Cap Group 13'   14='Soc Cap Group 14'    15='Soc Cap Group 15'    16='Soc Cap Group 16'
17='Soc Cap Group 17'   18='Soc Cap Group 18'    19='Soc Cap Group 19'    20='Soc Cap Group 20'
21='Soc Cap Group 21'   22='Soc Cap Group 22'    23='Soc Cap Group 23'    24='Soc Cap Group 24'
25='Soc Cap Group 25'   26='Soc Cap Group 26'    27='Soc Cap Group 27'    28='Soc Cap Group 28'
29='Soc Cap Group 29'   30='Soc Cap Group 30'    31='Soc Cap Group 31'    32='Soc Cap Group 32'
33='Soc Cap Group 33'   34='Soc Cap Group 34'    35='Soc Cap Group 35'    36='Soc Cap Group 36'
37='Soc Cap Group 37'   38='Soc Cap Group 38'    39='Soc Cap Group 39'    40='Soc Cap Group 40'
 .='Missing Soc Cap'
```

DWELLING TYPE FORMATS

Variable : H_DwgTp **Dwelling Type (detailed)** **1981**

Format : f81dtyp

```
 1='Private Dwelling: Separate house (1 H/H)'
 2='Private Dwelling: House or flat attached to business premises'
 3='Private Dwelling: House (2 or more H/Hs) with shared facilities'
 4='Private Dwelling: House with other private dwellings attached'
 6='Private Dwelling: Self-contained flat or apartment'
 7='Private Dwelling: Bach, Crib, hut (not in work camp)'
 8='Private Dwelling: Mobile or moveable home'
 9='Other private dwellings, incl. temporary'
10='Hotel, Motel, Private Hotel, Guest House'
11='Boarding or Rooming House'
12='Educational Institution (school hostel etc)'
13='Religious institution (convent, monastery)'
14='Hospital, convalescent home'
15='Home for Elderly'
16='Welfare Inst. (church hostel, night shelter)'
17='Other camp or hostel (youth or immigration)'
18='Motor camp'
19='Prison, police lock up or station'
20='Armed forces camp, vessel etc'
21='Staff quarters, nurses home etc'
22='Seasonal group quarters (shearers etc)'
23='Vessel (not Navy)'
24='Communes'
25='Other non private dwelling (fire stations etc)'
```

Variable : H_DwgTp **Dwelling Record Type** **1996**

Format : f96dtyp

```
 1='Private Dwelling'          2='Non-Private Dwelling'
```

Variable : H_DwgTpG **Dwelling Type** **1986,1991,1996**

Format : fdtyp

```
 1='Permanent/Fixed'          2='Semi-Permanent'              3='Temporary/Mobile'
 4='Hospitals'                5='RestHome for the Elderly'    8='Other Private Dwellings'
 9='Others'                   .='Missing DwellType'
```

NATURE OF OCCUPANCY FORMATS

Variable : H_Noccy **Nature of Occupancy** **1981,1986**

Format : f81Occ

```
  1='Owned with Mortgage'
  2='Owned without Mortgage'
  3='Rented/Leased from Priv. Person/Comp.-Furnished'
  4='Rented/Leased from Priv. Person/Comp.-Unfurnished'
  5='Rented/Leased from Priv. Person/Comp.-Furnishing Not Spec.'
  6='Rented/Leased from Housing Corp.'
  7='Rented/Leased from Other Govt Depts'
  8='Rented/Leased from Local Authority'
  9='Rented/Leased from Landlord Not Spec.'
 10='Provided Rent Free'
.,99='Tenure Not Specified'
```

Variable : H_Noccy **Nature of Occupancy** **1991**

Format : f91Occ

```
0='Owned with Mortgage'                          1='Owned without Mortgage'
2='Provided Rent Free'                           3='Private Person (rented or leased)'
4='Real Estate Agency (rented or leased)'        5='Housing Corporation (rented or leased)'
6='Other Government Departments (rented or leased)'  7='Local Authority (rented or leased)'
8='Landlord Not Specified (rented or leased)'    .,9='Tenure Not Specified'
```

Variable : H_Noccy **Nature of Occupancy** **1996**

Format : f96Occ

```
  1='Owned with Mortgage'
  2='Owned without Mortgage'
  3='Owned, Mortgage not specified'
  4='Provided Rent Free'
  5='Private Person (rented or leased)'
  7='Housing New Zealand (rented or leased)'
  8='Other Central Government Agency (rented or leased)'
  6='Local Authority or City Council (rented or leased)'
  9='Business, Real Estate Agency or other organisation (rented or leased)'
 10='Landlord Not Specified (rented or leased)'
 11='Not Owned, Rental status not specified'
.,99='Tenure Not Specified'
```

Variable : H_Tenure **Tenure** **1996**

Format : f96Tenr

```
  1='Owned with Mortgage'              2='Owned without Mortgage'
  3='Owned, Mortgage Not Specified'    4='Provided Rent Free'
  5='Rented'                           6='Not Owned, Rental Status Not Specified'
 .,9='Not Specified'
```

HOUSEHOLD TYPE FORMAT

Variable : H_Type **Household Type** **1981**

Format : f81HHT

```
10='1F(C)-Husband & wife only (no absentees)'
11='1F(C)-Husband & wife+unmarr. children (no absentees)'
12='1F(C)-Husband & wife only (no children absent, other person(s) absent)'
13='1F(C)-Husband & wife+unmarr. children (no children absent, other person(s) absent)'
20='1F(I)-Husband & wife only (>=1 children absent)'
21='1F(I)-Husband & wife+unmarr. children (1 or more children absent)'
22='1F(I)-Husband & wife only (>=1 children absent & other person(s) absent)'
23='1F(I)-Husband & wife+unmarr. children (>=1 children absent & other person(s) absent)'
24='1F(I)-One parent+unmarr. children (spouse temp. absent)'
25='1F(I)-One parent+unmarr. children (no absentees)'
26='1F(I)-One parent+unmarr. children (>=1 children and spouse temp. absent)'
27='1F(I)-One parent+unmarr. children (>=1 children absent, spouse perm. absent)'
28='1F(I)-One parent+unmarr. children (>=1 children, spouse & other persons temp. absent)'
29='1F(I)-One parent+unmarr. children (>=1Child&other persons temp. absent, spouse perm. Absnt)'
30='1F(I)-One parent+unmarr. children (no children absent, spouse & other persons temp. absent)'
31='1F(I)-One parent+unmarr. children (no children absent,oth.pers temp.absent,spse prm. Absnt)'
40='1F+OP-Husband & wife+other person(s) (w/wo absentees)'
41='1F+OP-Husband & wife, unmarr. children + other person(s) (w/wo absentees)'
42='1F+OP-1Par,unmar.child+oth.pers related to par (w/wo child & oth.pers. abs,spouse tmp. abs)'
43='1F+OP-1Par,unmar.child+oth.pers related to par(w/wo child & oth.pers.abs,spouse perm.abs)'
44='1F+OP-1Par,unmar.child+oth.pers not related to par(w/wo child &oth.pers.abs,spouse tmp.abs)'
45='1F+OP-1Par,unmar.child+oth.pers not related to par(w/wo child &oth.pers.abs,spouse prm.abs)'
46='1F+OP-1Par,unmar.child+oth.pers reld&not reld to par(w/wo child&oth.pers.abs,spse tmp.abs)'
47='1F+OP-1Par,unmar.child+oth.pers reld&not reld to par(w/wo child&oth.pers.abs,spse prm.abs)'
50='2F-1stFam:Hus&wife w/wo unmarr. children (no abs);2ndFam:Hus&wife w/wo unmarr.child(no abs)'
51='2F-1stFam:Hus&wife w/wo unmarr. children (no abs);2ndFam:One parent+unmarr.children(no abs)'
52='2F-1stFam:One parent+unmarr. children (no abs);2ndFam:Hus&wife w/wo unmarr.children(no abs)'
53='2F-1stFam:One parent+unmarr. children (no abs);2ndFam:One parent+unmarr. children (no abs)'
54='2F-Two familes (with absentees)'
55='2F-Two families+other person(s) (w/wo absentees)'
60='3F-Three or more families, w/wo other person(s) (no absentees)'
61='3F-Three or more families, w/wo other person(s) (with absentees)'
70='NF-Relatives only'
71='NF-Persons not related'
72='NF-Related and non-related persons'
80='1P-Usually a one-person household (no absentees)'
81='1P-One-person household (related person(s) temp. absent)'
82='1P-One-person household (non-related person(s) temp. absent)'
83='1P-One-person household (related & non-related persons temp. absent)'
84='Not elsewhere classified'
 .='Missing'
```

USUAL HOUSEHOLD COMPOSITION FORMAT

Variable : H_UsHHC **Usual Household Composition** **1981**

Format : f81UHC

```
 1='Couples Only'                                        2='Couples with Children'
 3='One Parent Family'                                   4='Couples Only plus Others'
 5='Couples with Children plus Others'                   6='One Parent Family plus Others'
 7='Two 2 Parent Families with or without Children'      8='Two Parent plus One Parent Family'
 9='Two 1 Parent Families'                              10='Three or More Families'
11='Non-Family Households'                              12='One-Person Households'
13='Not Elsewhere Classified. i.e. Visitors only'      .,99='Missing or Not Applicable'
```

Variable : H_UsHHC **Usual Household Composition** **1986,1996**

Format : fhhc

```
  1='HH with children with Sole Parent'          2='HH with children not Sole Parent'
  3='Sole Person Household'                      9='Other Groupings'
```

USUAL RESIDENCE FORMATS

Variable : UsInd **Usual Residence Indicator** **1981**

Format : f81USI

```
  0='Different from Census Night Address'          1='Same as Census Night Address'
  2='Different from CN but Meshblock Same'         .='Not Applicable'
```

Variable : UsInd **Usual Residence Indicator** **1986,1991,1996**
Variable : UsInd91 **Usual Residence Indicator 1991** **1991**

Format : fUSI

```
  1='Same as Census Night Address'     2='Elsewhere in New Zealand'      4='No Fixed Abode'
  5='Not Specified (within NZ)'        .='Not Applicable'
```

Variable : YrsUR **Years at Usual Residence** **1986**

Note: Values in data-set are single values, not grouped

Format : f86YUR

```
    0='Less than 1 year'          1='One Year'              2- 4=' 2- 4 Years'
 5- 9=' 5- 9 Years'            10-19='10-19 Years'        20-29='20-29 Years'
30-39='30-39 Years'            40-49='40-49 Years'        50-59='50-59 Years'
60-69='60-69 Years'            70-79='70-79 Years'           97='80 Years or More'
   99='Not Specified'              .='Not Applicable'
```

Variable : YrsUR **Years at Usual Residence** **1991**

Note: Values in data-set are single values, not grouped

Format : f91YUR

```
    0='Less than 1 year'          1='  One Year'            2- 4=' 2- 4 Years'
 5- 9=' 5- 9 Years'            10-19='10-19 Years'        20-29='20-29 Years'
30-39='30-39 Years'            40-49='40-49 Years'        50-59='50-59 Years'
60-69='60-69 Years'            70-79='70-79 Years'        80-89='80-89 Years'
90-96='90-96 Years'               97='97 Years or More'      98='Not Specified (>=5 years)'
   99='Not Specified'              .='Not Applicable'
```

Variable : YrsUR **Years at Usual Residence** **1996**

Note: Values in data-set are single values, not grouped

Format : f96YUR

```
    0='Less than 1 year'          1='  One Year'            2- 4=' 2- 4 Years'
 5- 9=' 5- 9 Years'            10-19='10-19 Years'        20-29='20-29 Years'
30-39='30-39 Years'            40-49='40-49 Years'        50-59='50-59 Years'
60-69='60-69 Years'            70-79='70-79 Years'        80-89='80-89 Years'
90-96='90-96 Years'               97='97 Years or More'      98='Unidentifiable'
   99='Not Specified'              .='Not Applicable'
```

TELEPHONE FORMAT

Variable : H_Teleph | **Telephone in Dwelling** | **1996**

Format : fTele

```
  1='Yes-Have Telephone'          2='No -No Telephone'          3='Unidentified'
 .,9='Not Specified'
```

INDUSTRY FORMATS

Variable : IndAnz1 | **ANZSIC Industry (1 xter)** | **1996**

Format : $fANZ

```
'A'='Agriculture, Forestry & Fishing'          'B'='Mining'
'C'='Manufacturing'                            'D'='Electricity, Gas & Water Supply'
'E'='Construction'                             'F'='Wholesale Trade'
'G'='Retail Trade'                             'H'='Accommodation, Cafes & Restaurants'
'I'='Transport & Storage'                      'J'='Communication Services'
'K'='Finance & Insurance'                      'L'='Property & Business Services'
'M'='Government Administration & Defence'      'N'='Education'
'O'='Health & Community Services'              'P'='Cultural & Recreational Services'
'Q'='Personal & Other Services'                  .='Missing'
```

Variable : Industry | **Industry Code (1 Digit)** | **1981,1991,1996**

Format : f1Ind

```
1='Agriculture, Hunting, Forestry & Fishing'   2='Mining & Quarrying'
3='Manufacturing'                              4='Electricity, Gas & Water'
5='Construction'                               6='Wholesale, Retail Trade & Restaurants&Hotels'
7='Transport, Storage & Communication'         8='Business & Financial Services'
9='Community, Social & Personal Services'    .,0='Missing or Not Adequately Defined'
```

Variable : Industry | **Industry Code (2 Digit)** | **1986**

Format : f2Ind

```
11='Agriculture & Hunting'                     12='Forestry & Logging'
13='Fishing'                                   21='Coal Mining'
22='Crude Petroleum & Natural Gas Production'  23='Metal Ore Mining'
29='Other Mining & Quarrying'                  31='Food, Beverage, Tobacco'
32='Textile, Apparel & Leathergoods'           33='Wood Processing & Wood Product Manufacture'
34='Manufacturing of Paper & Paper Products; Printing & Publishing'
35='Manufacture of Chemicals & of Chemical,Petroleum,Coal,Rubber & Plastic Materials'
36='Concrete,Clay,Glass,Plaster,Masonry,Asbestos & Related Mineral Product Manufacture'
37='Basic Metal Industries'
38='Manufacture of Fabricated Metal Products,Machinery & Equipment'
39='Other Manufacturing Industries'            41='Electry,Gas & Steam'
42='Water Works & Supply'                      51='Construction of Buildings'
52='Construction other than Buildings'         53='Ancillary Construction Services'
61='Wholesale Trade'                           62='Retail Trade'
63='Restaurants & Hotels'                      71='Transport & Storage'
72='Communication'                             81='Financing'
82='Insurance'                                 83='Real Estate & Business Services'
91='Public Administration & Defence'           92='Sanitary & Cleaning Services'
93='Social & Related Community Services'       94='Recreational & Cultural Services'
95='Personal & Household Services'           .,0='Missing or Not Defined Adequately'
```

Variable : O_Occ2X | **Occupation Code - 2 Digits Occ68** | **1981,1986,1991,1996**

Variable : O_OccSp2X | **Spouse Occupation Code - 2 Digits Occ68** | **1986**

Format : f2xOcc

```
 1='Physical Scientists(1)+Statisticians,Mathematicians,Systems Analysts(8)'
 2='Architects, Engineers (2)'
 3='Architects, Engineers (3)'
 4='Aircraft and Ships Officers'
 5='Life Scientists'
```

```
 6='Medical, Dental, Veterinary (6)'
 7='Medical, Dental, Veterinary (7)'
11='Accountants(11)+Economists(9)'
12='Jurists'
13='Teachers'
14='Workers in Religion'
15='Authors, Journalists & Related Writers'
16='Sculptors, Painters, Photographers & Related Creative'
17='Composers&Performing Artists(17)+Athletes,Sportsmen/Sportswomen(18)'
19='Professional, Technical'
21='Managers(21)+Legislative Officials&Government Administive(20)'
30='Clerical Supervisors'
31='Government Executive Officials'
32='Stenog. Typists, Card & Tape Punching Machine Operators'
33='Bookkeepers, Cashiers(33)+Computing Machine Operators(34)'
35='Transpost & Communications Supervisors'
37='Mail Distribution Clerks(37)+Transport Conductors(36)'
38='Telephone and Telegraph Operators'
39='Clerical nec'
40='Managers (Wholesale & Retail Trade)'
41='Working Proprietor (Wholesale & Retail Trade)'
42='Sales Supervisors and Buyers'
43='Technical Salesperson, Representative & Manufacturing'
44='Insurance, Real Estate, Securities Salespersons'
45='Salespersons, Shop Assistants(45)+Sales Workers nec(49)'
50='Managers (Catering & Lodging Services)'
51='Working Proprietors (Catering & Lodging Services)'
53='Cooks, Waiters, Waitresses, Bartenders etc.'
54='House Staff&Housekeeping Services nec(54)+Housekeeping&Related Service Supervisors(52)'
55='Building Caretakers, Charworkers'
56='Launderers, Drycleaners & Pressers'
57='Hairdresser, Barber, Beautician etc.'
58='Protective Service Workers - Including Armed Forces'
59='Service Workers nec'
60='Farm Managers & Supervisors'
61='Farmers'
62='Agriculture & Animal Husbandry Workers'
63='Forestry Workers'
64='Fishermen, Hunters'
70='Production Supervisors & General Foremen/Women'
71='Mine/Quarrymen, Well-Drillers etc.'
72='Metal Processors'
73='Wood Preparation Workers & Paper Makers'
74='Chemical Processors'
75='Spinners,Weavers,Knitters(75)+Tanners,Fellmongers&Pelt Dressers(76)'
77='Food&Beverage Processors(77)+Tobacco Preparers&Tobacco Product Makers(78)'
79='Tailors etc. Upholsterers'
80='Shoemakers & Leather Goods Maker'
81='Cabinet Makers & Related Woodworkers'
83='Blacksmith,Toolmakers,Machine Tool Operators(83)+Stone Cutters&Carvers(82)'
84='Machinery Fitters, Assemblers etc. Not Electrical'
85='Electrical Fitters etc.Electronic Wrkrs(85)+Broadcast,Sound-EqupOper&Cine.Projectists(86)'
87='Plumbers,Welders,Sheet & Structural Metal Preparers & Erectors'
88='Jewellery & Precious Metal Workers'
89='Glass Forgers, Potters'
90='Rubber & Plastics Product Makers'
92='Printers(92)+Paper & Paperboard Product Makers(91)'
93='Painters'
94='Production'
95='Bricklayers, Carpenters & Other Construction'
96='Stationary Engine & Related Equipment Operator'
97='Material, Dockets & Freight Handlers etc.'
```

```
98='Transport Equipment Operators'
99='Labourers nec'
 .='Missing'
```

OCCUPATIONAL CLASS and SEI FORMATS

Elley Irving from 3 digit NZSCO68 codes according to Neil Pearces concordance

Variable : Occ68 used to form O_Elleyirv — **1981,1986,1991,1996**

InFormat : $i3Einp

```
'011','012','013','021','022','023','024','025','026','027','051','052','053','061','063'=1
'065','075','081','082','090','110','121','122','129','131','132','139','191','192','195'=1
'201','202'=1
'028','029','031','041','042','043','067','069','076','079','083','084','133','134','135'=2
'141','151','159','179','193','194','199','211','212','219','300','310','441'=2
'014','032','033','034','035','036','037','038','039','054','062','064','066','068','073'=3
'077','149','162','163','171','172','173','174','175','180','321','322','331','339','342'=3
'351','352','359','392','393','394','395','399','400','421','422','431','432','442','443'=3
'500','510','581','582','592','611','612','613','614','615','616','701','702','703','704'=3
'705','706','707','708','709','733','734','832','844','852','861','961'=3
'071','072','074','161','341','360','370','380','391','410','451','452','490','531','583'=4
'589','591','600','617','619','641','713','753','762','775','776','777','797','811','819'=4
'820','841','842','843','846','847','848','849','851','854','855','856','857','859','862'=4
'871','874','880','893','902','921','922','923','924','925','926','929','941','951','952'=4
'953','954','955','956','959','969','972','973','981','982','983','989'=4
'520','532','540','560','570','599','621','628','631','632','649','721','722','723','724'=5
'725','726','727','728','729','735','741','742','743','744','745','749','751','752','754'=5
'755','756','759','761','771','772','773','774','778','781','782','783','789','791','792'=5
'793','794','795','796','799','812','831','833','834','835','839','845','853','872','873'=5
'891','892','894','895','899','901','910','927','931','939','943','957','971','974','979'=5
'984','985'=5
'551','552','622','623','624','625','626','627','629','711','712','731','732','779','801'=6
'802','803','942','949','986','990','991'=6
'996','998','999','997'=9
.=.
other=99
```

Variable : O_Elleyirv — **Elley-Irving Index** — **1981,1986,1991,1996**

Variable : O_ElleyirvSp — **Elley-Irving Index (Spouse)** — **1986**

Format : fEI

```
  1='EI Class 1'                    2='EI Class 2'                    3='EI Class 3'
  4='EI Class 4'                    5='EI Class 5'                    6='EI Class 6'
  7='EI Class 7 (Farmers)'       .,9,99='EI Class 9 (Miss or NS)'
```

EGP from 4 digit NZSCO68 codes

Variable : Occ68 (4 digit) used to form O_EGP — **1981,1986,1991,1996**

InFormat : $i4EGP

```
'0110','0120','0131','0132','0134','0135','0139','0211','0212','0219','0221','0222','0223'=1
'0224','0225','0229','0231','0232','0233','0239','0241','0242','0243','0244','0245','0249'=1
'0250','0260','0270','0281','0289','0291','0292','0293','0294','0299','0411','0412','0413'=1
'0421','0422','0423','0424','0425','0429','0430','0511','0512','0513','0519','0521','0522'=1
'0523','0524','0525','0529','0531','0532','0533','0534','0535','0536','0539','0611','0613'=1
'0614','0615','0617','0619','0631','0651','0652','0659','0670','0810','0821','0822','0823'=1
'0901','0902','0909','1101','1102','1103','1104','1109','1211','1219','1221','1290','1311'=1
'1312','1391','1392','1740','1921','1929','2011','2012','2019','2021','2022','2029','2111'=1
'2119'=1
```

```
'0141','0311','0312','0313','0314','0315','0319','0321','0322','0323','0324','0325','0326'=2
'0327','0329','0331','0332','0333','0334','0339','0341','0342','0349','0350','0360','0370'=2
'0380','0390','0541','0542','0620','0641','0649','0661','0680','0690','0711','0712','0713'=2
'0714','0715','0716','0719','0721','0722','0731','0740','0750','0761','0762','0763','0771'=2
'0779','0791','0792','0793','0794','0795','0799','0830','0841','0849','1321','1322','1329'=2
'1331','1332','1341','1349','1350','1399','1411','1412','1419','1490','1510','1591','1592'=2
'1593','1594','1599','1610','1621','1622','1623','1624','1625','1629','1631','1632','1633'=2
'1639','1711','1712','1713','1714','1719','1721','1722','1731','1732','1739','1750','1791'=2
'1799','1801','1802','1803','1804','1809','1911','1912'=2
'1913','1919','1931','1933','1939','1941','1943','1949','1950','1990','2121','2129','2191'=2
'2192','2193','2199','3001','3009','3101','3102','3109','3510','3520','3591','4001','4002'=2
'4009','4211','4219','4221','4222','4223','4224','4229','4310','4411','4412','4419','4420'=2
'4431','4436','4439','5001','5002','5003','5004','5009','5822','5823','5824','5829','5891'=2
'3211','3212','3213','3214','3215','3216','3219','3220','3311','3312','3313','3314','3315'=3
'3316','3319','3391','3392','3393','3394','3395','3399','3411','3412','3421','3422','3429'=3
'3592','3593','3594','3595','3596','3597','3599','3601','3602','3603','3609','3801','3802'=3
'3803','3804','3809','3911','3912','3913','3919','3920','3931','3932','3933','3934','3935'=3
'3939','3941','3942','3943','3944','3949','3951','3952','3991','3992','3993','3994','3999'=3
'4321','4322','4511','4512','4513','4514','4515','4516','4517','4519','4521','4523','4524'=3
'4529','4900','5201','5202','5204','5205','5209','5321','5322','5323','5329','5911','5912'=3
'5919','5920','5991','5994','8621','8622','8629'=3
'4101','4102','4109','5101','5103','5104','5105','5109'=4
'6000','6111','6119','6121','6122','6129','6131','6132','6133','6134','6139','6141','6142'=6
'6143','6144','6145','6149','6151','6152','6159','6160','6171','6172','6173','6174','6179'=6
'6191','6192','6199'=6
'7010','7020','7030','7040','7050','7060','7070','7091','7092','7093','7094','7095','7099'=7
'5311','5312','5313','5319','5701','5702','5703','5704','5705','5709','5811','5812','5819'=8
'7111','7112','7113','7114','7119','7131','7132','7133','7134','7135','7139','7211','7219'=8
'7221','7229','7231','7239','7241','7242','7249','7251','7252','7259','7260','7271','7272'=8
'7273','7279','7281','7282','7289','7311','7312','7319','7320','7321','7322','7329','7350'=8
'7531','7532','7533','7539','7541','7542','7543','7544','7545','7546','7547','7549','7551'=8
'7559','7561','7562','7564','7569','7611','7612','7613','7614','7615','7619','7621','7622'=8
'7629','7720','7731','7732','7733','7734','7736','7739','7741','7742','7743','7744','7745'=8
'7749','7761','7762','7763','7764','7765','7769','7771','7772','7779','7781','7782','7783'=8
'7784','7785','7786','7789','7911','7912','7919','7921','7922','7929','7931','7932','7939'=8
'7941','7942','7943','7944','7949','8011','8012','8013','8019','8021','8031','8032','8039'=8
'8110','8191','8192','8193','8194','8195','8196','8199','8201','8209','8311','8312','8313'=8
'8319','8321','8322','8323','8329','8331','8332','8339','8351','8352','8353','8359','8391'=8
'8392','8393','8394','8395','8396','8397','8398','8399','8414','8415','8416','8419','8421'=8
'8422','8423','8424','8425','8426','8427','8429','8431','8432','8433','8439','8440','8461'=8
'8462','8463','8465','8466','8469','8472','8492','8499','8511','8512','8513','8514','8515'=8
'8519','8521','8522','8529','8540','8551','8552','8553','8554','8555','8559','8560','8571'=8
'8572','8573','8574','8575','8579','8591','8599','8610','8711','8713','8714','8719','8721'=8
'8722','8723','8729','8731','8733','8734','8736','8738','8739','8741','8742','8743','8749'=8
'8801','8802','8803','8804','8805','8809','8911','8912','8913','8914','8915','8916','8917'=8
'8919','8921','8922','8923','8924','8929','8931','8932','8933','8939','9211','9212','9213'=8
'9214','9216','9219','9222','9223','9224','9225','9229','9230','9240','9250','9261','9262'=8
'9269','9270','9291','9299','9311','9312','9313','9319','9411','9412','9419','9431','9432'=8
'9439','9511','9512','9513','9519','9521','9522','9529','9541','9542','9543','9549','9551'=8
'9552','9559','9591','9592','9593','9594','9595','9599','9720','9731','9732','9733','9739'=8
'9741','9742','9743','9744','9745','9746','9747','9748','9749'=8
'3701','3702','3703','3704','3709','5401','5403','5404','5405','5409','5510','5521','5522'=9
'5523','5529','5601','5602','5603','5604','5609','5892','5893','5894','5895','5899','5992'=9
'5995','5996','5997','5998','5999','7120','7291','7299','7324','7330','7340','7410','7420'=9
'7431','7432','7439','7440','7450','7491','7492','7493','7499','7512','7515','7519','7521'=9
'7522','7529','7591','7592','7711','7719','7751','7752','7753','7754','7755','7756','7759'=9
'7791','7792','7793','7799','7811','7819','7820','7830','7890','7951','7952','7953','7959'=9
'7961','7962','7963','7969','7991','7992','7999','8022','8023','8024','8029','8121','8123'=9
'8129','8341','8342','8343','8349','8464','8471','8481','8482','8483','8489','8491','8531'=9
'8532','8533','8534','8539','8940','8951','8952','8953','8959','8991','8992','8993','8999'=9
'9011','9012','9013','9014','9015','9016','9019','9021','9022','9023','9029','9101','9102'=9
'9103','9104','9109','9391','9393','9399','9421','9422','9429','9491','9493','9494','9499'=9
```

```
'9531','9532','9539','9560','9570','9611','9612','9613','9614','9615','9616','9617','9618'=9
'9619','9691','9692','9693','9694','9695','9696','9699','9711','9712','9713','9714','9715'=9
'9716','9717','9718','9719','9791','9792','9799','9810','9821','9822','9829','9831','9832'=9
'9839','9841','9842','9843','9849','9851','9852','9853','9854','9856','9859','9860','9891'=9
'9892','9899','9900','9901','9902','9903','9904','9905','9906','9907','9908','9909','9910'=9
'9911','9912','9913','9914','9915','9916','9917','9918'=9
'6211','6219','6221','6222','6229','6230','6241','6242','6243','6244','6245','6247','6248'=10
'6249','6250','6260','6271','6272','6273','6279','6281','6282','6283','6284','6289','6291'=10
'6292','6299','6311','6312','6313','6314','6315','6316','6317','6319','6321','6322','6323'=10
'6324','6325','6326','6327','6329','6411','6412','6413','6414','6415','6419','6491','6492'=10
'6493','6494','6495','6496','6497','6499'=10
.=99
other=0
```

Variable : O_EGP — **EGP** — **1981,1986,1991,1996**

Variable : O_EGPSp — **EGP (Spouse)** — **1986**

Format : fEGP

```
 1='High grade professionals & administrators'
 2='Lower grade professionals & higher grade technicians'
 3='Routine non-manual employees, sales & service workers'
 4='Small proprietors & administrators with employees'
 5='Small proprietors & administrators without employees'
 6='Farmers & self-employed fishermen'
 7='Lower grade technicians & foremen'
 8='Skilled manual workers'
 9='Semi & unskilled manual workers'
10='Agricultural workers'
 0='Missing'
```

Formats to assign SEI by three digit NZSCO90 codes

Variable : NZSCO90 used to create O_SEI91 — **1991,1996**

InFormat : $i90SEI

```
'611'=22.4   '826'=22.7   '523'=22.9   '612'=25.1   '512'=26.7   '911'=27.3   '513'=29.4
'915'=29.8   '813'=29.8   '744'=30.2   '521'=32.9   '514'=33.7   '822'=33.7   '914'=34.4
'828'=35.7   '823'=36.4   '743'=36.4   '913'=36.5   '827'=37.5   '422'=37.7   '824'=37.9
'821'=38.2   '741'=38.3   '613'=38.6   '832'=38.7   '245'=38.8   '825'=39.0   '833'=39.6
'829'=39.6   '614'=39.8   '742'=40.4   '414'=40.8   '412'=42.0   '912'=42.2   '421'=43.1
'812'=43.4   '731'=44.2   '841'=44.2   '413'=44.3   '411'=44.7   '711'=45.2   '721'=45.5
'522'=45.8   '811'=46.7   '733'=48.8   '712'=48.7   '723'=48.7   '713'=49.2   '834'=49.4
'814'=49.5   '732'=49.9   '334'=50.4   '336'=50.3   '323'=51.3   '815'=51.3   '511'=52.5
'724'=53.2   '722'=53.5   '313'=53.8   '122'=54.0   '011'=54.2   '335'=54.9   '312'=55.3
'322'=55.5   '234'=56.6   '331'=56.6   '321'=58.3   '338'=58.8   '223'=60.1   '332'=60.1
'816'=60.2   '515'=61.2   '831'=61.3   '233'=61.5   '243'=61.8   '315'=62.2   '114'=62.6
'311'=63.4   '333'=64.7   '121'=65.0   '241'=70.9   '214'=73.2   '314'=73.5   '213'=74.8
'244'=75.3   '232'=76.6   '235'=77.0   '231'=77.6   '221'=79.2   '211'=81.9   '112'=82.0
'212'=82.6   '111'=83.9   '242'=88.9   '113'=89.8   '222'=90.0   '337'=10.0   other=.
```

1996 NZSEI scores from 3 digit NZSCO95 code

Variable : NZSCO95 used to create O_SEI96 — **1996**

InFormat : $i96SEI

```
'111'=63   '112'=69   '113'=90   '114'=46   '121'=69   '122'=50   '211'=68
'212'=71   '213'=60   '214'=56   '221'=58   '222'=89   '223'=45   '231'=69
'232'=61   '233'=43   '234'=45   '235'=58   '241'=61   '242'=83   '243'=44
'244'=62   '245'=32   '311'=46   '312'=47   '313'=46   '314'=65   '315'=44
'321'=45   '322'=45   '323'=33   '331'=48   '332'=51   '333'=46   '334'=29
'335'=40   '336'=49   '337'=31   '338'=42   '411'=33   '412'=34   '413'=28
'414'=30   '421'=30   '422'=27   '511'=38   '512'=18   '513'=19   '514'=20
'515'=44   '521'=22   '522'=30   '523'=64   '611'=22   '612'=34   '613'=31
'614'=37   '711'=36   '712'=32   '713'=43   '721'=32   '722'=39   '723'=34
```

```
'724'=37    '731'=34    '732'=27    '733'=35    '741'=26    '742'=28    '743'=26
'744'=20    '811'=36    '812'=27    '813'=19    '814'=30    '815'=39    '816'=48
'821'=25    '822'=23    '823'=23    '824'=26    '825'=30    '826'=10    '827'=24
'828'=24    '829'=24    '831'=46    '832'=26    '833'=27    '834'=32    '841'=29
'911'=21    '912'=32    '913'=25    '914'=19    '915'=18    other=.
```

Variable : Used to group SEI91 into classes		1991,1996

InFormat : i91sei

```
75-90=1     60-<75=2     50-<60=3     40-<50=4     30-<40=5     10-<30=6     other=9
```

Variable : O_SEI91v	SEI 91 Values	1991,1996
Variable : O_SEI91vFa	SEI 91 Values (Father)	1991
Variable : O_SEI91vMo	SEI 91 Values (Mother)	1991
Variable : O_SEI91vPr	SEI 91 Values (Parent)	1991

Note: Values in data-set are single values, not grouped

Format : f91sei

```
75 -  90='NZSEI Class 1'          60 -< 75='NZSEI Class 2'          50 -< 60='NZSEI Class 3'
40 -< 50='NZSEI Class 4'          30 -< 40='NZSEI Class 5'          10 -< 30='NZSEI Class 6'
   other='NZSEI Class 9 (Miss or NS)'
```

Variable : Used to group SEI96 into classes	1996

InFormat : i96sei

```
66-90=1     56-65=2     42-55=3     32-41=4     24-31=5     10-23=6     other=9
```

Variable : O_SEI96v	SEI 96 Values	1996

Note: Values in data-set are single values, not grouped

Format : f96sei

```
66-90='NZSEI Class 1'               56-65='NZSEI Class 2'               42-55='NZSEI Class 3'
32-41='NZSEI Class 4'               24-31='NZSEI Class 5'               10-23='NZSEI Class 6'
other='NZSEI Class 9 (Miss or NS)'
```

Variable : If SEI class variable formed, this would be its format	1991,1996

Format : fnsei

```
   1='NZSEI Class 1'              2='NZSEI Class 2'                3='NZSEI Class 3'
 4,8='NZSEI Class 4'              5='NZSEI Class 5'                6='NZSEI Class 6'
   7='NZSEI Class 7 (Farmers)'    9='NZSEI Class 9 (Miss or NS)'
```

Farmer Flag from 4 digit NZSCO68

Variable : Used on 4 digit NZSCO68 to group into farmers and non-farmers	1981,1986,1991,1996

InFormat : $i68Frm

```
'6111','6121','6122','6129','6131','6132','6133','6134','6141','6142','6143','6144','6145'=1
'6149','6151','6152','6159','6160','6171','6172','6173','6174','6179','6191','6192','6211'=1
'6219','6221','6222','6229','6230','6241','6242','6243','6244','6245','6246','6248','6249'=1
'6250','6260','6271','6272','6273','6279','6289','6291','6292','6299','7511','7783','7789'=1
'7799','9919','0532','0662','6119','6139'=1
other=0
```

Variable : O_FarmFlg	Farmers Occupation Flag	1981,1986,1991,1996
Variable : O_FarmFlgSp	Farmers Occupation Flag (Spouse)	1986
Variable : O_FarmFlgFa	Farmers Occupation Flag (Father)	1991
Variable : O_FarmFlgMo	Farmers Occupation Flag (Mother)	1991
Variable : O_FarmFlgPr	Farmers Occupation Flag (Parent)	1991

Format : fFarmF

```
   0='Not a Farmer'          1='Farmer'
```

RELIGION FORMATS

Variable : Religion **Religion - Main Groups** **1981**
Variable : Religion **Religion - Treat Groups With Caution** **1986,1996**

Format : f81relg

```
 1='Anglican Church'                                  2='Presbyterian Church of New Zealand'
 3='Roman Catholic'                                   4='Methodist'
 5='Christian N.O.D.'                                 6='Baptist'
 7='Church of Jesus Christ of Latter Day Saints'      8='Ratana Establishment Church of NZ'
 9='Protestant N.O.D.'                               10='Brethren'
11='Salvation Army'                                  12='Jehovah''s Witness'
13='Seventh Day Adventist'                           96='Other Religions'
97='No Religion'                                     98='Object'
99='Not Specified'
```

Variable : Used to group 1986 religion variable to be consistent with 1981 **1986**
Note: SNZ is currently investigating what groupings to use over time

InFormat : i86rlg

```
 1= 1     2= 2     3= 3     4= 4    17= 5     5= 6     96= 7     97= 8      9= 9
10=10    11=11    12=12    13=13     6=97     8=98     94,999,.=99       other=96
```

Variable : Used to group 1996 religion variable to be consistent with 1981 **1996**
Note: SNZ is currently investigating what groupings to use over time

InFormat : $i96rlg

```
'2031'= 1   '2271'= 2   '2090'= 3   '2201'= 4   '2100'= 5   '2050'= 6   '2171'= 7
'2141'= 8   '2290'= 9   '2070'=10   '2311'=11   '2151'=12   '2012'=13   '7599'=96
'8051'=97   '8091'=98   '8071','8111','9999',.=99
```

SMOKING FORMATS

Variable : SmkCur — **Current Smoking Status** — **1981**

Format : fSmkC

```
  0='Never Smoked Cigarettes'        1='Used to Smoke'       2='Currently Smoking Regularly'
.,9='Not Specified'
```

Variable : SmkEver — **Ever Smoked** — **1996**

Format : fSmkE

```
1='Yes - Smoked'       2='No - Never Smoked'       3='Inidentifiable'      .,9='Not Specified'
```

Variable : SmkQnt — **Quantity of Cigarettes Smoked on 23 Mar 1981** — **1981**

Note: Values in data-set are single values, not grouped

Format : fSmkQ

```
 0='Nil, but otherwise smoked regularly'    1- 4=' 1- 4 Cigarettes'     5- 9=' 5- 9 Cigarettes'
10='   10 Cigarettes'                      11-14='11-14 Cigarettes'    15-19='15-19 Cigarettes'
20='   20 Cigarettes'                      21-24='21-24 Cigarettes'    25-29='25-29 Cigarettes'
30='   30 Cigarettes'                      31-34='31-34 Cigarettes'    35-39='35-39 Cigarettes'
40='   40 Cigarettes'                      41-44='41-44 Cigarettes'    45-49='45-49 Cigarettes'
50='   50 Cigarettes'                      51-54='51-54 Cigarettes'    55-59='55-59 Cigarettes'
60='   60 Cigarettes'                      61-64='61-64 Cigarettes'    65-69='65-69 Cigarettes'
70='   70 Cigarettes'                      71-74='71-74 Cigarettes'    75-79='75-79 Cigarettes'
80='   80 Cigarettes'                      81-84='81-84 Cigarettes'    85-89='85-89 Cigarettes'
90='   90 Cigarettes'                      91-94='91-94 Cigarettes'    95-96='95-96 Cigarettes'
97='97 or more Cigarettes'                    98='Not Applicable'       .,99='Not Specified'
```

Variable : SmkReg — **Smoking Regularly** — **1996**

Format : fSmkR

```
1='Smoking Regularly'     2='Not Smoking Regularly'    3='Unidentifiable'    9='Not Specified'
```

Variable : SmkStat — **Smoking Status** — **1996**

Format : fSmkS

```
  1='Smoker'                  2='Ex-Smoker'            3='Never Smoked Regularly'
  4='Unidentifiable'         .,9='Not Specified'
```

GENERIC FORMAT

Variable : AU1Yr — **Same Area Unit of Residence 1 Year Ago** — **1981**

Format : fYesNo

```
  0='No'       1='Yes'       .='Missing'
```

GENERAL NUMBER COUNT FORMATS

Variable : H_Mveh — **Number of Motor Vehicles in H/H** — **1996**

Format : f3num

```
0='Nil'          1='1'          2='2'          3='3 or more'      .,9='Not Specified'
```

Variable : H_Mveh — **Number of Motor Vehicles in H/H** — **1986,1991**
Variable : H_NabTot — **Total Number of Absentees in H/H** — **1996**

Format : f5num

```
0='Nil'          1='1'          2='2'          3='3'              4='4'
5='5 or more'  .,9,99='Not Specified'
```

Variable : H_FtJob — **Number of Full-time Jobs in H/H** — **1986**

Format : f7num

```
0='Nil'          1='1'          2='2'          3='3'              4='4'
5='5'            6='6'          7='7 or more'   .,8,9='Not Specified'
```

Variable : H_PtJob — **Number of Part-time Jobs in H/H** — **1986**

Format : f7gnum

```
0='Nil'          1='1'          2='2'          3='3 or 4'         5='5'
6='6'            7='7 or more'   .,8,9='Not Specified'
```

Variable : H_Bcars — **Number of Business Cars in H/H** — **1981**
Variable : H_Bdrms — **Number of Bedrooms** — **1986,1991**
Variable : H_Mveh — **Number of Private Cars in H/H** — **1981**
Variable : H_Nadult — **Number of Adults aged 20+ in H/H (on C/N)** — **1981**
Variable : H_Nadult — **Number of Adults aged 16+ in H/H (on C/N)** — **1986**
Variable : H_NChn — **Number of Children aged 0-15 in H/H (on C/N)** — **1981,1986**
Variable : H_Pbike — **Number of Pushbikes in H/H** — **1981**

Format : f8num

```
0='Nil'          1='1'          2='2'          3='3'              4='4'
5='5'            6='6'          7='7'          8='8 or more'      .,9='Not Specified'
```

Variable : H_NabCh — **Number of Children Absent in H/H** — **1981**
Variable : H_NabTot — **Total Number of Absentees in H/H** — **1981**

Format : f9num

```
0='Nil'          1='1'          2='2'          3='3'              4='4'
5='5'            6='6'          7='7'          8='8'              9='9 or more'
.='Not Applicable'
```

Variable : H_Bdrms — **Number of Bedrooms** — **1996**

Format : f14num

```
1='1'            2='2'          3='3'          4='4'              5='5'
6='6'            7='7'          8='8'          9='9'             10='10'
11='11'         12='12'        13='13'        14='14 or more'    98='Unidentifiable'
.,99='Not Specified'
```

Variable : H_Bdrms — **Number of Bedrooms** — **1981**
Variable : H_PerFam — **Number of People in Family** — **1996**

Format : f20num

```
1='1'            2='2'          3='3'          4='4'              5='5'
6='6'            7='7'          8='8'          9='9'             10='10'
11='11'         12='12'        13='13'        14='14'            15='15'
16='16'         17='17'        18='18'        19='19'            20='20 or more'
.,99='Not Specified'
```

Variable : H_OccTot — **Total Number of Occupants in H/H** — **1996**

Note: Values in data-set are single numbers, not grouped

Format : f500nm

```
1- 49='  1- 49'      50- 99=' 50- 99'     100-199='100-199'     200-299='200-299'
300-399='300-399'    400-499='400-499'    500-998='500 or more'
```

```
.,999='Missing'
```

ABSENTEE FORMAT

Variable : AbsentFlg — **Absentee Indicator** — **1981,1986,1991,1996**

Format : fabs

```
0='Non-Absentee'      1='Absentee'
```

Variable : PerType — **Personal Record Type** — **1981,1986,1991,1996**

Format : fPRecT

```
1='Absentee'       3='NZ Adult'       4='NZ Child'
```

IMPUTATION FIELD FORMATS

Variable : Imp — **Imputation Indicator** — **1991**

Format : f91limp

```
0='None'                        1='Age'                          2='Sex'
3='Total Hours Worked'          4='Sex & Age'                    5='Sex & Total Hours Worked'
6='Age & Total Hours Worked'    7='Sex, Age & Total Hours Worked'  .='Not Applicable'
```

Variable : ImpAge — **Age Imputation Indicator** — **1991**

Format : f91lage

```
1='Absentee Age Imputation Code 1'   2='Absentee Age Imputation Code 2'      .='Not Applicable'
```

Variable : ImpAge — **Age Imputation Indicator** — **1996**

Format : f96lage

```
 .,0='No Imputation'                 1='Imputed from Family'              2='No Information'
   3='From Dwelling Form'            4='Conflicting Information'          5='Unknown Code 5'
   6='Unknown Code 6'                7='Unknown Code 7'                   8='Unknown Code 8'
   9='Unknown Code 9'               10='Unknown Code G'                  11='Unknown Code X'
```

Variable : ImpForm — **Form Imputated Indicator (Dummy Form)** — **1996**

Format : f96ldum

```
   .='Record Present'              1='Dummy Record Code 1'            2='Dummy Record Code 2'
```

Variable : ImpLFS — **Imputation in Labour Force Status** — **1996**

Format : f96lLFS

```
   .='No Imputation'                      1='Any Value Imputed'
   2='Full or Part Time Imputed'          3='Unemployed or Not in Labour Force'
```

Variable : ImpRes — **Imputation in Usual Residence Status** — **1996**

Format : f96lres

```
  1='Possibly Area Unit Known'        2='Possibly TLA Known'        3='Regional Council Known'
  4='Possibly No Information'         .='No Imputation'
```

Variable : ImpSex — **Imputation in Sex** — **1996**

Format : f96lsex

```
.,0='No Imputation Done'      1='Imputed from Name or Relationship'    2='Stochastic Imputation'
```

Variable : ImpMonth — **Month of Age Imputation Indicator** — **1981,1986,1991,1996**

Format : flMth

```
   0='No Imputation'              1='Age in Months Imputed'
```

LINKING OF MORTALITY RECORDS FORMAT

Variable : Link | **Matched** | **1981,1986,1991,1996**

Format : flink

```
  0='Not Linked'          1='Linked'
```

CAUSE OF DEATH FORMATS

Variable : Used on ICDA to form ICD_Gp | **1981,1986,1991,1996**

InFormat : $iicd

```
' '                                  = '   '
'001'-'139XX','320'-'323XX','390'-'392XX'= '001'   (Communicable Diseases)
'460'-'466XX','590'-'590XX','595'-'595XX'= '001'
'614'-'616XX','680'-'686XX','711'-'711XX'= '001'
'771'-'771XX'                        = '001'
'140'-'152XX','155'-'161XX','163'-'173XX'= '140'   (Other Cancer)
'175'-'184XX','186'-'209XX'          = '140'
'153'-'154XX'                        = '153'   (Colorectal Cancer)
'162'-'162XX'                        = '162'   (Lung/Bronchus Cancer)
'174'-'174XX'                        = '174'   (Breast Cancer (female))
'185'-'185XX'                        = '185'   (Prostate Cancer)
'250'-'250XX'                        = '250'   (Diabetes)
'393'-'399XX','402'-'402XX'          = '390'   (Other Heart Disease)
'404'-'409XX','415'-'429XX'          = '390'
'400'-'401XX','403'-'403XX','440'-'459XX'= '400'   (Other Cardiovascular Disease)
'410'-'414XX'                        = '410'   (IHD)
'430'-'438XX'                        = '430'   (Cerebrovascular Disease)
'470'-'478XX','494'-'494XX','497'-'519XX'= '470'   (Other Respiratory)
'480'-'487XX'                        = '480'   (Pnuemonia/influenza)
'490'-'492XX','495'-'496XX'          = '490'   (COPD)
'493'-'493XX'                        = '493'   (Asthma)
'740'-'759XX'                        = '740'   (Congenital)
'760'-'770XX','772'-'779XX'          = '760'   (Perinatal)
'798'-'79809'                        = '798'   (SIDS)
'800'-'809XX','826'-'949XX'          = '800'   (Unintentional Injury other than RTC)
'810'-'825XX'                        = '810'   (RTC)
'950'-'959XX','980'-'989XX'          = '950'   (Suicide)
'960'-'979XX','990'-'999XX'          = '960'   (Violent)
                                 other = '999'
```

Variable : ICD_Gp | **International Cause of Death (ICD)** | **1981,1986,1991,1996**

Format : $ficd

```
'001'='Communicable Diseases'              '153'='Colorectal Cancer'
'162'='Lung/Bronchus Cancer'               '174'='Breast Cancer'
'185'='Prostate Cancer'                    '140'='Other Cancer'
'250'='Diabetes'                           '410'='IHD'
'390'='Other Heart Disease'                '430'='Cerebrovascular Disease'
'400'='Other Cardiovascular Disease'       '480'='Pnuemonia/Influenza'
'490'='COPD'                               '493'='Asthma'
'470'='Other Respiratory'                  '740'='Congenital'
'760'='Perinatal'                          '798'='SIDS'
'810'='RTC'                                '950'='Suicide'
'960'='Violent'                            '800'='Unintentional Injury other than RTC'
'999'='Other Causes'                       '   '='Not Dead/Linked'
```

Variable : CauseDeath	Cause of Death (4 groups)	1981,1986,1991,1996

Format : f4dth

```
1='Cancer'    2='CVD'    3='Injury inc Sui&Int'    4='Other Causes'    .='Not Dead/Linked'
```

SEASON OF DEATH

Variable : SeasDth	Season at Death	1981,1986,1991,1996

Format : fseason

```
1='Summer'    2='Autumn'    3='Winter'    4='Spring'    .='Missing'
```

HOSPITALISATION FORMATS

Variable : PostAUIn	Post Census Area Unit Indicator	1991,1996

Format : fPostC

```
0='Not Hospitalised Post-Census'    1='Hospitalised Post-Census'    .='Not Applicable'
```

Variable : PreAUIn	Pre Census Area Unit Indicator	1991,1996

Format : fPreC

```
0='Not Hospitalised Pre-Census'    1='Hospitalised Pre-Census'    .='Not Applicable'
```

DISABILITY FORMATS

Variable : DisCode	Long-Term Disability or Handicap	1996

Format : fDisCd

```
1='Have Disability'    2='No Disability'    .,9='Not Specified'
```

Variable : DisInd	Disability Indicator (from HealthProb & DisCode)	1996

Format : fDisIn

```
0='No Disability Indicated'    1='Disability Indicated'    .,9='Not Specified'
```

HEALTH PROBLEMS FORMATS

Variable : HealthProb	Health Problems	1996

Format : fHProb

```
0='No Specified Health Problems'    1='Specified Health Problems'    .,9='Not Specified'
```

Variable : HealthProb_A	Health Problem 1	1996
Variable : HealthProb_B	Health Problem 2	1996
Variable : HealthProb_C	Health Problem 3	1996

Format : fHProbD

```
1='Had difficulties with everyday activities that people your age can usually do'
2='Had difficulties with communicating, mixing with others or socialising'
3='Had difficulties with any other activity that people your age can usually do'
.,9='Did not have difficulties doing task'
```

5.4. Duplicate records

5.4.1 Comparing Duplicates

Check if values for NHI data are the same on both records. The variables checked were: date of birth; sex; ethnicity; and domicile code. (Because of the way duplicates were identified by NZHIS, all these values were the same for possible duplicate records.) If there were no differences in the above four checks, then we could say that the records were a **true duplicate.** If there were two or more differences, we classed the records as **not duplicates.** If there was one difference, we classed the records as **possible duplicates** requiring manual review. Manual review of these records was often able to identify what were probably minor miscodings in a true duplicate pair.

5.4.2 Amalgamating Duplicate Records

For those records classed as true duplicates, data was amalgamated into a single mortality record (filling in as many variables as possible).

1) If there were two fields in the data, one from the NHI source, and one from the NMDS source, and if any of the data differed, we put one value in the NHI field, and the other in the NMDS field.
2) We merged comment fields to make sure we kept all the information, separated by ']' if they differed.
3) If there was data in both of the ICDA, ICDB1, ICDB2 or ICDC fields (medical classifications of causes of death), we tried to keep the more specific values so used either the first record or the values that did not end in a '9' or an 'X'.
4) If the occupation value differed we used the one that was not '999' or if they were both not '999', then the one that did not start with '99', otherwise we used the first value.
5) For all the data that we decided were duplicates, one record had a registration district that started with a '9', the other did not. We used the record with the non -'9' value for Registration District for obtaining the Registration District, Registration Year, Registration Quarter, Registration Entry Number, Id number and Place of Death values.
   On almost all (or possibly all) of these records, the place of death differed between the two records. We used the one that was associated with the non-'9' Registration District field.
6) The records that we no longer wanted were given the value of 'D' in the Duplicate Status variable and at the final step in creation of the dataset for Record Linkage, this field was deleted.
7) The other records in this group of duplicate sets were given the value 'Y' for Duplicate, or 'P' for Possible Duplicates for the Duplicate Status variable. We wanted to not loose the information that some records had differed in one key field.