

Glossary

General terms

- Differential health selection** (See glossary definition of **health selection** first.)
Health selection that occurs when a socioeconomic mortality gradient is assessed among the active labour force only, due to exclusion from the active labour force that is differential by that socioeconomic factor.
Classically, differential health selection has been described for occupational class mortality gradients. Here the gradient is underestimated when only *current* occupation is available for the assignation of occupational class. This underestimation is because the lower occupational classes (based on *usual* occupation) are more likely to be forced out of the labour force than the higher occupational classes when sick, causing the observed mortality risk/rate among the lower occupational classes (based on *current* occupation) to be underestimated.
- Drift health selection** (See glossary definition of **health selection** first.)
Health selection that occurs when people drift down the socioeconomic ladder consequent on their health status. For example, poor health may both lower one's (current) income and be associated with an increased risk of death, causing an overestimate of the association of (usual) income with mortality.
- Health selection** '... the artificial raising or lowering of the average health of people with a particular characteristic associated with the process by which that characteristic is acquired or lost. The mortality of a population with that characteristic is affected by health-related mobility if the health of people acquiring or losing the characteristic differs systematically from others with the characteristic.' (Fox et al 1987).
In this report, it is crucial to consider two types:
- **Drift health selection** (see glossary definition)
 - **Differential health selection** (see glossary definition).
- It is also useful to consider health selection (either of the two types above) as occurring over the short or long term. With three years of follow-up in the NZCMS, it was only possible to investigate short-term health selection.
- Linkage bias** The biases by demographic and socioeconomic factors in the proportion of mortality records linked to a census record.
- Multivariate analysis** In this report, regression analyses that include more than one socioeconomic factor as independent variables.

Univariate analysis

'It is increasingly common in the medical literature to use the term *univariate analysis* to refer to analyses which examine only a single explanatory variable's relationship to a [single] response [outcome] variable.' (Armitage and Colton 1998, p.4663). In this report, age and ethnicity is treated as fundamental demographic variables that must be adjusted for. Therefore, 'univariate analysis' in this report refers to the association of one socioeconomic factor (eg, income) with mortality, controlling for age and ethnicity. (All analyses were conducted separately by sex.) (See **multivariate analysis**.)

Glossary of record linkage terms

Agreement frequency ratio or agreement odds	The agreement frequency ratio is the odds of a particular matching variable agreeing among links versus agreeing among non-links , that is the m probability divided by the u probability.
Agreement weight	In record linkage using Automatch [®] , the weight is the logarithm (base two) of the agreement frequency ratio .
Automatch [®]	Commercial probabilistic record linkage software used in the NZCMS.
Blocking	A procedure used in record linkage to reduce the number of possible comparisons. That is, the records on both files are divided into blocks (eg, area of residence), and record linkage is conducted within these blocks only.
Blocking variable	Variable used to 'block' files in record linkage. In the NZCMS, the blocking variables were geocodes: [meshblock] and [census area units].
Combined frequency ratio or combined odds	For any given comparison pair , the product of the agreement frequency ratios for each matching variable that agrees and the disagreement frequency ratios for each matching variable that disagrees. The combined frequency ratio constitutes the information on which the overall 'betting odds' in favour of (or against) a correct match are based, and hence pairs categorised as links or non-links.
Combined weight	For any given comparison pair , the sum of the agreement weights for each matching variable that agrees and the disagreement weights for each matching variable that disagrees. The combined weight constitutes the information on the relative weight in favour of (or against) a correct match, and hence pairs are categorised as links or non-links.
Disagreement frequency ratio or disagreement odds	The disagreement frequency ratio is the odds of a particular matching variable disagreeing among links versus disagreeing among non-links , that is [1 minus the m probability] divided by [1 minus the u probability].

Disagreement weight	In record linkage using Automatch® , the weight is the logarithm (base two) of the disagreement frequency ratio .
False link or false positive	As conceptualised in the linkage bias analysis in the NZCMS, those census respondents who did not die but were linked.
False non-link or false negative	As conceptualised in the linkage bias analysis in the NZCMS, those census respondents who did die but were not linked.
Files	The sets of records to be compared in the record linkage – in the NZCMS the mortality and census files.
Frequency ratio	(See ‘agreement frequency ratio’ and ‘disagreement frequency ratio’.)
Linkage bias	As operationalised in the NZCMS, the <i>misclassification bias of the mortality outcome due to the record linkage</i> . Taking all census records as the population, one can create a two-by-two table of links (equivalent to the ‘diagnostic test’ for vital status) by actual vital status. It is then possible to consider the sensitivity, specificity, positive predictive value and negative predictive value of the record linkage for vital status.
Links	Those comparison pairs that are categorised during the record linkage as <i>probably</i> including the same person’s mortality and census record. As people die only once, each mortality record can only be linked to one census record. In most record linkage projects (including the NZCMS) it is impossible to determine which links are matches or non-matches , although it is assumed that the majority (hopefully the vast majority) of links are matches.
m probability	The probability that a matching variable agrees given that the comparison pair being examined is categorised as a link . It may be global (eg, one common <i>m</i> probability for all values of [year of birth]) or value specific (eg, different <i>m</i> probabilities for each value of [year of birth]).
Match	Following Newcombe (1988), a pair of mortality and census records that are for the same person (ie, the pair is ‘correct’). It is usually impossible to verify in a record linkage project whether any pair is indeed a match – rather, the relative probability of a pair being a match is estimated.
Matching variable	Variable that is common to both the mortality and census files, and hence available for comparing pairs of records during the record linkage. In the NZCMS, matching variables were [day of birth], [month of birth], [year of birth], [sex], [ethnicity], [country of birth].
Match-run	The full series of passes that make up the record linkage.
Non-links	Those comparison pairs that are categorised during the record linkage as <i>probably not</i> including the same person’s mortality and census record.

Non-match	Following Newcombe (1988), a pair of mortality and census records that are <i>not</i> for the same person (ie, the pair is 'incorrect').
Pair or comparison pair	Any comparison of a pair of records from different files . The theoretical number of pairs is the product of the number of records on the two files (eg, if there were 100 mortality records and 10,000 census records, the number of possible pairs is 1 million). Blocking dramatically reduces the number of possible pairs.
Pass	A given specification of matching variables, blocking variable(s), mortality and census records to be processed, <i>m</i> and <i>u</i> probabilities, and other parameters set by the operator of Automatch ®. A sequence of passes (up to eight in Automatch®) makes up a match-run .
Positive predictive value	As conceptualised in the linkage bias analysis in the NZCMS, the proportion of the linked census records who did die.
Records	In the NZCMS, the separate mortality events on the mortality file and the separate census entries of the census file.
Sensitivity	As conceptualised in the linkage bias analysis in the NZCMS, the proportion of the census cohort who did die and who were linked during the record linkage.
Specificity	As conceptualised in the linkage bias analysis in the NZCMS, the proportion of the census cohort who did not die and who were not linked during the record linkage.
True link or true positive	As conceptualised in the linkage bias analysis in the NZCMS, those census respondents who died and who were linked.
True non-link or true negative	As conceptualised in the linkage bias analysis in the NZCMS, those census respondents who did not die and who were not linked.
<i>u</i> probability	The probability that a matching variable agrees given that the comparison pair being examined is categorised as a non-link (ie, the probability that variables agree purely by chance among non-links). It may be global (eg, one common <i>u</i> probability for all values of [year of birth]) or value specific (eg, different <i>u</i> probabilities for each value of [year of birth]).
Weight	See agreement weight and disagreement weight .