Public Health Monograph Series
No. 13
ISSN 1178-7139

LINKAGE OF CENSUS AND CANCER REGISTRATIONS, 1981-2004

CancerTrends Technical Report: Number 1

June Atkinson
Caroline Shaw
Tony Blakely
James Stanley
Kate Sloane

March 2010

A technical report published by the Department of Public Health,
University of Otago, Wellington

ISBN 978-0-473-16586-4

Acknowledgements

Development of CancerTrends commenced in 2003. Many people have been involved in the inception and progress of this project. We have tried to acknowledge them all here, but inevitably we will have omitted names and we apologise for that.

Initial funding for CancerTrends to test whether the linkage was feasible was provided by the Ministry of Health. Additionally without other work done by the Ministry of Health to clean up the addresses on the Cancer Register the project would have been unlikely to succeed. CancerTrends is now funded by the Health Research Council of New Zealand and the Ministry of Health.

At the Ministry of Health we thank staff of (the former) Public Health Intelligence and New Zealand Health Information Services particularly Martin Tobias, Chris Lewis, Susan Hanna, Tracey Vandenberg, Barry Borman (former manager Public Health Intelligence).

Also, thanks to staff at SNZ who have championed CancerTrends – Paul Brown and the Government Statistician, and those who have laboured over the linkage and other work involved particularly Andy Smith, Barb Lash, Nalin Patel, Lilian Morrison, Anapapa Mulitalo, Steve White, John Upfold, Brian Cosgriff, Judith Archibald, Shari Mason, Chris Hansen and Jackie Dixon.

Finally thanks to all the authors of the previous New Zealand Census-Mortality Study technical reports, of whose work we have borrowed from considerably.

Executive Summary

CancerTrends is a record linkage study of the 1981, 1986, 1991 1996 and 2001 censuses, each to 5 years of subsequent cancer registrations (nearly 4 in the case of 2001), creating five short duration cohort studies of the entire New Zealand population followed up for cancer incidence. The study follows the precedent of the New Zealand Census-Mortality Study (NZCMS) in method and overall aims. The goals of CancerTrends are to describe ethnic and socioeconomic trends in cancer incidence, correct for (any) bias in the recording of ethnicity between cancer and census data, and answer specific research questions (e.g. smoking and given cancers).

Anonymous and probabilistic record linkage was used to link census and cancer data. 73.2% (1981-86 cohort) to 81.7% (2001-04 cohort) of eligible cancer registrants were linked to census data. Of these links, 95.2%% (1981-86) to 96.9% (2001-04) were estimated to be true links. The percentage of eligible cancer registrants linked varied by socio-demographics, with lower rate of linkage for males, 15-24 year olds (53% to 65% across cohorts), Māori (61% (males 1981-86) to 82% (females 2001-04)), and Pacific (63% (males 1981-86) to 82% (females 2001-04). This lower linkage success, if associated with variables to be used as exposures in future cohort analyses (e.g. ethnicity, socioeconomic position), will result in differential misclassification bias of the mortality outcome. To prevent such bias, the linked census-cancer pairs were weighted up to be representative of all eligible cancer registrants, using linkage weights.

Linkage weights were calculated by first using logistic regression to determine independent predictors of linkage success. The results from this analysis where then used to determine an automated algorithm for aggregation of strata of all eligible census registrants (for each cohort) by sex, ethnicity, deprivation, age, cancer diagnosis, territorial authority, residential mobility of area unit, and time since census. This algorithm aggregated ‘adjacent’ cells with at least 75% of eligible cancer records linked, or five or more links were in each strata. The linked pairs in these strata were then assigned an inverse probability weight, which was merged with the main cohort file.

A similar algorithm was used to assign weights to the ‘highly probable links’ (i.e. where ethnicity was not influential), and ‘unlock ratios’ calculated with these weights as in the

NZCMS. In the early 1980s there was an undercount of nearly a third for Māori cancer registrants (for cancer registry ethnicity compared to census ethnicity), decreasing to a 15% undercount in 2001-04. The undercount for Pacific similarly decreased over time to 10%. Undercounting of Asian cancer registrants fell from 68% to 13%.

Much of this technical report details the data management of the linked census-cancer files, including variable generation and formats, to underpin future cohort analyses. Of note, we also used multiple imputation to impute missing income, education and ethnicity data, with the aim of reducing possible selection bias in future cohort analyses. This had not previously been undertaken with NZCMS data. It consumed considerable resources to undertake given the large file sizes. Unfortunately, as demonstrated within this report, the imputation outputs were implausible. For example, the number of imputed high income Māori was implausible.

Some recommendations arise from this technical report:

1. The 2006 census should be linked to mortality and cancer data simultaneously, thus merging the NZCMS and CancerTrends projects and data files. This will resulted in a large saving of resources.
2. Methodological research to understand why the multiple imputation failed is warranted, with a view to recommendations for the 2006-2011 cohorts (and retrospectively previous cohorts).
3. At least one more linkage of cancer and mortality data to the census is warranted for monitoring of ethnic and socioeconomic inequalities in New Zealand, and smoking contributions given the 2006 census includes smoking data.
4. The reliability of using Health datasets only, with the NHI as a pseudo-population register, for monitoring mortality and cancer inequalities into the future requires active consideration and feasibility research – preferably in parallel with the 2006 census linkage to assess validity.

Statistics New Zealand Security Statement

CancerTrends was initiated by Professor Tony Blakely and his co-researchers from the University of Otago, Wellington. It was approved by the Government Statistician as a Data Laboratory project under the Microdata Access Protocols. All research publications are based on researcher initiated ideas.

Access to the data used in this study was provided by Statistics New Zealand under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975. The results presented in this study are the work of the author, not Statistics New Zealand

(The full security statement is available at http://www.uow.otago.ac.nz/nzcms-info.html)

Table of Contents

List of Tables

List of Figures

Glossary

Area unit	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 cancer records contain two different dates of birth for the same individual)
AutoMatch®	The original version of the software package for carrying out probabilistic record linkage. The latest version is called QualityStage™
Bias analysis	Estimating any systematic differences between linked and unlinked cancer 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.
Cancer Register	A population based cancer register for all newly disgnosed malignant diseases.
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-cancer cohort datasets to determine differences in cancer rates by social factors.
Comparison pair	Any possible comparison of a record from one file with a record from another file. In CancerTrends comparison pairs consist of one census and one cancer 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., QualityStage™ has found two census (A) records that match the same cancer (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' cancer record from a duplicate pair – i.e., QualityStage™ has found two cancer (B) records that match the same census (A) record. One of these cancer records will be listed as part of a matching pair (MP), and the other as a duplicate match (DB).
Datalab	Statistics New Zealand secure Data Laboratory where researchers work on their projects.

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).
Extent of disease	The Cancer Register variable that approximates clinical stage at diagnosis.
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 CancerTrends, File A refers to census records, while File B refers to cancer 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.
International Classification of Disease (ICD)	The world standard classification system of disease and procedures. The current version is ICD-10.
Linkage bias	Systematic differences by socio-demographic factors (e.g. age, deprivation) between linked and unlinked cancer records.
Links	A comparison pair that is accepted as being highly likely to apply to the same individual. In CancerTrends > 95% of links are matches or true links.
MP pair	A linked (probably matching) pair of records, consisting of one census record (A) and one cancer record (B). The total weight for the pair is above the specified cut-off for the given QualityStage™ 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.
Mortality collection	A mortality dataset which classifies the underlying cause of death for all deaths registered in New Zealand.
National Health Index	An NZHIS dataset, containing data for nearly every individual in New Zealand with a unique identifier (the NHI number) and demographic variables. This data updated every time a person uses public health services (e.g. outpatient visits, diagnostic investigations). The NHI dataset can be linked to NMDS or cancer registration events for the same individual by means of the NHI number, allowing good linkage of datasets within the health sector.
National Minimum Data Set	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.
New Zealand Census-Mortality Study	The New Zealand Census-Mortality Study (NZCMS) now consists of five cohorts of anonymously, probabilitically linked Census and Mortality records. See http://www.uow.otago.ac.nz/nzcms-info.html.
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	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 rounding	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 is 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).
QualityStage™	Latest version of the software package for carrying our probabilistic record linkage. The original version used was AutoMatch®.
SAS	Software package used to do data manipulation and data management, as well as statistical analyses. SAS 8.2 used in Statistics New Zealand and SAS 9.1.3 used at University of Otago
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	In contect of record linkage, 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 (use area unit as the preferred name)
DHB	District Health Board (n=21)
HH	Household
HPL	Highly probable links
ICCC	International Classification of Childhood Cancer
MB	Meshblock
MOH	Ministry of Health
nMPA (or nonMPA)	Non-Maori/Pacific/Asian
NHI	National Health Index
NMDS	National Minimum Data Set
NZCMS	New Zealand Census-Mortality Study
NZDep	New Zealand Deprivation Index
NZSEI	New Zealand Socio-Economic Index (an occupational class index)
PPV	Positive predictive value
RHA	Regional Health Authority (n=4; health administrative boundaries used in 1990s)
SEER	Surveillance, Epidemiology and End Results. A source of cancer statistics in the United States. http://seer.cancer.gov/
SNZ	Statistics New Zealand
UOW	University of Otago, Wellington
WSM	Wellington School of Medicine (now called University of Otago, Wellington)

Purpose of this Report

CancerTrends consists of five cohort studies of the entire New Zealand population, linking the 1981, 1986, 1991, 1996 and 2001 censuses each to five years of subsequent NZ Cancer Registry data (3.83 years for 2001 census). These five cohorts were created retrospectively using anonymous and probabilistic record linkage. This report documents the technical processes used to 'create' the CancerTrends cohorts. The audience is intended to be those interested in the technical aspects of this study, and those who may use the datasets created for their own research.

The CancerTrends Study

This section covers the background and rationale for CancerTrends and an introduction to the methodology.

Introduction

CancerTrends is a Health Research Council funded research project with the following aims for the 2007-10 period.

- To determine ethnic and socio-economic trends in cancer incidence in NZ from 1981-2004.
- To determine ethnic and socio-economic trends in cancer survival in NZ.
- To answer a range of specific research questions:
    - Is the increasing colorectal cancer incidence among Pacific people evident for both overseas born and New Zealand-born Pacific people, recent immigrants? Are there cohort effects?
    - Are the increasing ethnic and socio-economic breast cancer mortality disparities due to trends in incidence, survival or both?
    - What is the contribution of active tobacco smoking to cancer incidence and trends?
    - What is the strength of association of passive smoking with lung and other cancers?
    - What are the trends in testicular cancer by ethnicity and socio-economic position in NZ?

CancerTrends' strategic importance is to provide quality epidemiological information for the second aim of the New Zealand Cancer Control Strategy, which is to decrease inequalities in cancer.

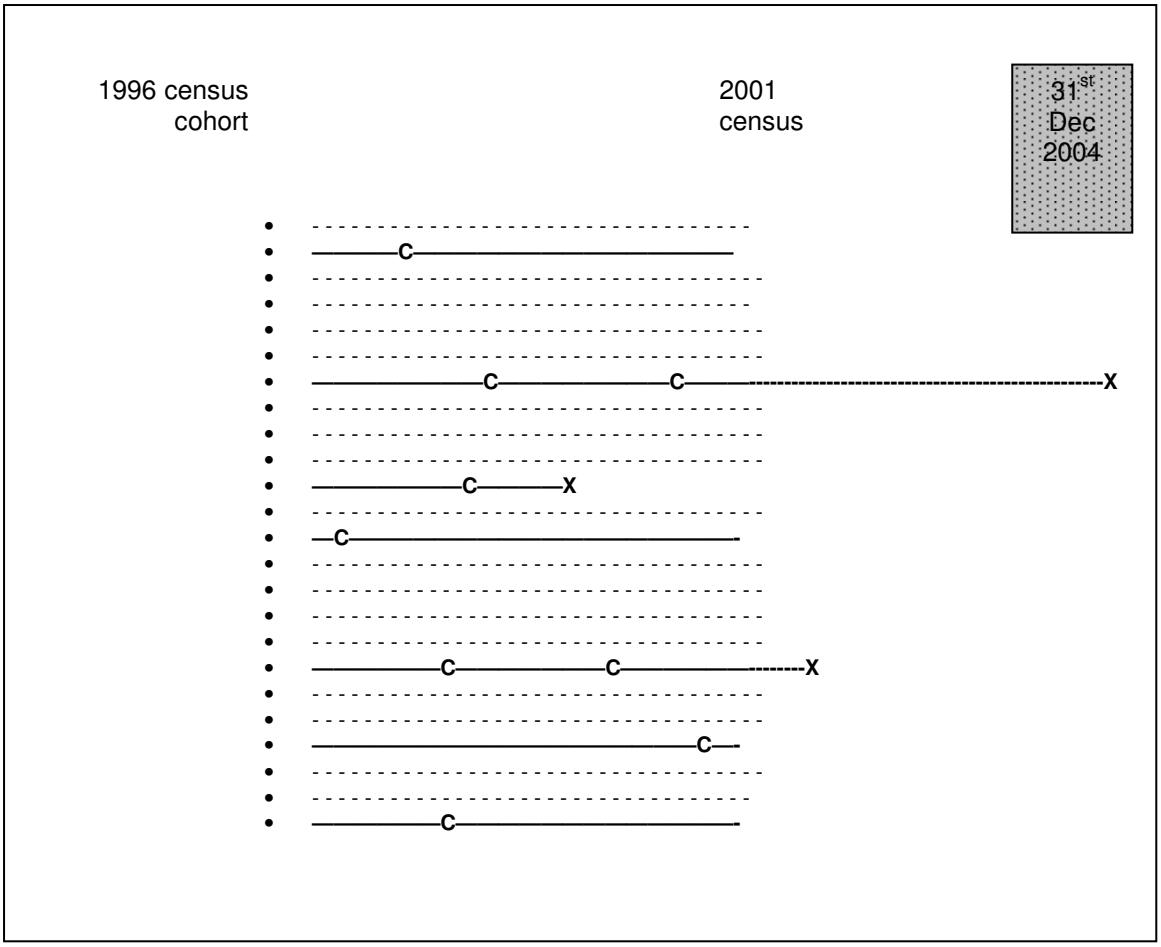
We also note that there will be many research questions that these datasets can contribute to answering.

Methods

Study Design

CancerTrends is five cohort studies of the entire New Zealand population. The cohorts were created retrospectively by linking two routinely collected datasets, census records and cancer registrations, using an anonymous and probabilistic linkage process. Cancer registrations in the five year period subsequent to the census (except 2004 in which only 3 years of data were available) were linked back to individual census records of the 1981, 1986, 1991, 1996 and 2001 censuses. Information on whether individuals on the cancer register were deceased was also available. This is illustrated for the 1996 cohort in Figure 1.

Figure 1 Illustration of 1996 CancerTrends cohort



C= incident cancer; X =death

The dots in Figure 1 signify individual census respondents, and the solid lines those individuals with at least one linked cancer event (c). In the absence of linkage to

migration data, and to other death data, censoring only occurs upon death (X) amongst those people with a linked cancer event.

Eligible participants for this study were New Zealand residents who either filled out a form or had one filled out for them on census night 1981, 1986, 1991, 1996, or 2001. Exposure information was obtained from the information on census forms. Cancer outcome was ascertained from the New Zealand Cancer Register (NZCR). Information on all individuals who met the study definition were obtained from NZHIS (see Table 1 for definition for each cohort).

Table 1 Cancer 'outcome' definition by cohort

Cohort	Inclusion criteria
1981-1986	Cancer diagnosed between 25 March 1981 and 4 March 1986 (date of diagnosis not date of notification) registered on NZ Cancer Registry and New Zealand resident. (Note : Five years of data).
1986-1991	Cancer diagnosed between 5 March 1986 and 5 March 1991 (date of diagnosis not date of notification) registered on NZ Cancer Registry and New Zealand resident. (Note : Five years of data).
1991-1996	Cancer diagnosed between 6 March 1991 and 5 March 1996 (date of diagnosis not date of notification) registered on NZ Cancer Registry and New Zealand resident. (Note : Five years of data).
1996-2001	Cancer diagnosed between 6 March 1996 and 6 March 2001 (date of diagnosis not date of notification) registered on NZ Cancer Registry and New Zealand resident. (Note : Five years of data).
2001-2004	Cancer diagnosed between 7 March 2001 and 31 December 2004 (date of diagnosis not date of notification) registered on NZ Cancer Registry and New Zealand resident. (Note : Three years ten months of data).

Note: individuals had to born on or before the previous census day to be in the cohort.

Data sources

This section looks briefly at the datasets that CancerTrends utilised, giving an overview of the Census of Population and Dwellings and the NZHIS datasets that information on individuals with cancer came from.

Census

The New Zealand Census of Population and Dwellings was initiated in 1851. Until 1956 the interval between each census varied between 3-10 years. Subsequent to

1956 the census has occurred every 5 years (Statistics New Zealand 2001). The Statistics Act 1975 mandates completion; every person in New Zealand on census day must complete an individual census form for the census or have one filled in for them. Post census enumeration surveys have been conducted since 1996 and indicate under-counting of between 1.6 and 2.2% (Statistics New Zealand 1996; Statistics New Zealand 2002; Statistics New Zealand 2006) although it is suspected that undercounting may actually be a bit greater (especially for young adults, males, and Mäori and Pacific).

The census collects mainly demographic and socio-economic information on individuals and households. Some additional questions are asked, for example smoking status in the 1981 and 1996 censuses. Copies of individual and household census questionnaires are available on www.stats.govt.nz . The specific census variables used in the anonymous linkage process are covered in Section 1.1.1 and those contained in the final datasets for analysis are in Table 43 in Appendix 1.

Health Datasets

Information on individuals who developed cancer was obtained from a number of different health datasets. This information was extracted by NZHIS and anonymised before it was delivered to the researchers at University of Otago. Information was needed from different datasets for the census linkage process, and for the analysis once the linkage was complete. Table 2 details the information taken from each of the datasets and the rest of the section describes each of the datasets in more detail.

Table 2 Information obtained from different NZHIS datasets

Dataset name	When	Information
National Health Index File	Information from time of data extraction- July 2004 for pilot 1996-2001 linkage, September 2006 and November 2007 for other cohorts	Date of birth Sex Date of death Ethnicity Meshblock and area unit of address
National Health Index 2001 data file	A one off copy of the NHI taken in December 2001	Date of birth Sex Ethnicity Meshblock and area unit of address
Cancer Register	Information from time cancer was registered	Date of birth Sex

		Country of birth Ethnicity Date of diagnosis Cancer type Cancer morphology Meshblock and area unit of address
Mortality Collection	Information from time of death	Date of birth Sex Country of birth Ethnicity Date of death Cause of death Meshblock and area unit of address
National Minimum Dataset Hospitalisation file	Interactions with health system, spanning 5 years before census to 5 years after census (individuals had differing numbers of interactions and thus differing amounts of information)	Date of birth (on later cohorts) Sex Country of birth (on later cohorts) Ethnicity Area unit of address Date of Admission Date of Discharge

NHI File

The National Health Index (NHI) is a unique number for an individual in relation to health events, such as hospitalisations and cancer registrations. The NHI file is a 'master' file that keeps the details on an individual's name, address, and ethnicity. This file is updated each time an individual has an interaction with tertiary health services and previous information is overwritten. This master file is then used as the 'link' between other datasets held by NZHIS so that a detailed profile of health need or interactions can be built up. Information obtained from the NHI file for CancerTrends is in Table 2. The National Health Index started in 1988.

Cancer Register

The New Zealand Cancer Register started in 1948, and is one of a number of international population based Cancer Registers. On 1 July 1994 the Cancer Registry Act 1993 and the Cancer Registry Act Regulations came into force, mandating that

all newly diagnosed malignant disease must be notified to the NZCR.[1] The Act and associated regulations defined, among other things, the scope of what was to be reported to the NZCR, the timeframes in which new cancers were to be reported and the manner in which they were to be reported. Importantly the Act mandated reporting by pathologists in laboratories. The Act and Regulations are available at www.nzhis.govt.nz . Note that benign neoplasms are not required to be reported to the NZCR.

A full description of the data collected by the NZCR is available in their data dictionary (New Zealand Health Information Service 2004). See Table 2 for brief details of the variables obtained for CancerTrends from this dataset.

Mortality Collection

The mortality dataset classifies the underlying cause of death for all deaths registered in New Zealand. All deaths in New Zealand are legally required to be registered with the Department of Internal Affairs. The Department releases this information to NZHIS. NZHIS obtain information on cause of death from a number of sources including: medical certificates of causes of death, Coroners' reports, electronic hospital discharge data from the National Minimum Dataset, private hospital discharge returns, the New Zealand Cancer Registry, the Department for Courts, the Police, the Land Transport Safety Authority, Water Safety NZ, Media Search, and from writing letters to certifying doctors, coroners, and medical records officers in public hospitals (New Zealand Health Information Service 2008).

Cause of death is currently coded in ICD-10-AM 2nd Edition, but deaths prior to 2000 are coded in ICD-9-CM-A.

The mortality database contains all deaths from 1988. NZHIS had spent a lot of effort trying to find any death records for earlier years of Cancer Registrations - however, it is incomplete, probably by about a third.

[1] For most of the history of the NZCR it was not compulsory to report cancers. Never-the-less case ascertainment, for some cancers at least, was thought to be relatively complete. For example colorectal cancer and testicular cancer. However, from the mid 1980s, due to changes in the health system structure and increasing societal concerns around patient privacy, the notification of new cases of cancer became problematic, with case ascertainment and case information declining. This was thought to be particularly problematic in cancers that did not require admission to hospital, such as early bowel cancer, melanoma and some cancers of the breast. (New Zealand Health Information Service 2000) Although, as far as we are aware, there was no formal assessment of the extent of under-reporting.

As a separate exercise NZHIS supplied us with mortality information (sex, date of birth, date of death, country of birth, ethnicity, cause of death and registration year) for all deaths registered between 1980 and 1990 (inclusive) and we used AutoMatch® to try and find any missed deaths. Because we did not have geocode information, whereas NZHIS had been able to use names, we were unable to be confident in finding any new mortality links. Thus, we do not recommend CacnerTrends data for registrations pre-1988 for use in survival analyses.

National Minimum Data Set

The National Minimum Dataset (NMDS) is a national collection of public and private hospital discharge information, including clinical information, for inpatients and day patients. Public hospital discharge information is available from 1988 onwards, and private hospital discharge information for publicly funded events has been collected since 1997 (New Zealand Health Information Service 2008).

Probabilistic record linkage summary

In order to create the CancerTrends datasets for analysis records from the census and the cancer register were linked to create five separate cohorts. Readers interested in the theory and approach to record linkage are directed to previous NZCMS publications (Fawcett, Atkinson et al. 2008; Hill, Atkinson et al. 2002) and technical details of the CancerTrends linkage are in a report of the process by Statistics New Zealand (Lash 2008). This chapter provides necessary details of the data used in the linkage, not included in the Statistics New Zealand report, and provides a summary table of the results.

1.1 Data used in 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 CancerTrends consisted of census records and cancer records. For each of the five census-cancer linkage projects the two files required different kinds of preparation in order to be suitable for QualityStage™ linkage. However, the final variables used for linkage were the same for all cohorts. This section focuses on the creation of census and cancer data files suitable to be linked.

1.1.1 Census linkage file

All census data is stored by Statistics New Zealand, and is kept under conditions of strict privacy. Since the data was not permitted to leave SNZ, SNZ undertook both the preparation of the census files, and the actual record linkage. The census data required for the CancerTrends cohorts were extracted from the census master-file for each of the 1981, 1986, 1991, 1996, and 2001 censuses and made into new files with the subset of variables that were needed for the linkage. The census variables that were used in the record linkage are presented in Table 1 and Table 3. Extra details of the derivation of some of variables used for the linkage is given below in the subsequent section (1.1.3)

Table 3 Census variables included for use in record linkage

Variable	Type	Len	Format	Label
AU	Char	6		Area Unit (Usual Residence Base 2001)
AgeC	Num	3	FAGE.	Age at census (years) – used for clerical review
Asian	Num	8	ETHA.	Ethnicity - Any Asian
AsianFix	Num	8	FIXA.	Asian - Ethnicity/Country of Birth Adjustment
BirthGp	Num	3	FBTHGP.	Country of Birth
DayC	Num	8		Day of Birth
ImpAge	Char	1	$FIMPAGE.	Age Imputation Indicator – used for clerical review
ImpSex	Num	8	FIMPSEX.	Sex Imputation Indicator – used for clerical review
MB	Char	7		Meshblock (Usual Residence Base 2001)
Maori	Num	8	ETHM.	Ethnicity - Any Maori i.e. Total Ethnicity
MonthC	Num	8	FMTH.	Month of Birth
PacFix	Num	8	FIXP.	Pacific - Ethnicity/Country of Birth Adjustment
Pacific	Num	8	ETHP.	Ethnicity – Any Pacific
Person_Id	Char	8 or 14		Census Person Id (length of 8 in 1996 and 2001, length of 14 in other years) – used for reference
YearC	Num	4		Year of Birth
nonMPA	Num	8	ETHO.	Ethnicity - Any non-Māori/Pacific/Asian (European/Other)
sex	Num	8	FSEX.	Sex

Source: (Lash 2008)

1.1.2 Health dataset variables used for linkage

The variables used from health datasets for the linkage process are detailed in Table 4. As mentioned above data were obtained from a number of different health datasets, including the NHI, mortality collection, NMDS, and the Cancer Register. This information was collated and arrays of options of the true value of specific variables were created. For example, up to five options (in order of likelihood of being correct) were used for day of month of birth (Day1W to Day5W).

As with previous NZCMS linkages, these multiple sources improve the linkage rate (refer to the various NZCMS technical reports for further information if needed).

Table 4 Health dataset variables used for linkage process

Variable	Type	Len	Format	Label
AsianFixW	Num	3	FIXA.	Asian Ethnicity born in Asian
AsianW	Num	3	ETHA.	Ethnicity - Any Asian on Cancer, Mort, NHI, Archived NHI files
BirthGpW	Num	3	FBTHGP.	Country of Birth Grouping
BirthGpW2	Num	3	FBTHGP.	Second Country of Birth Grouping – used for clerical review
Day1W	Num	3		Day of Birth Option 1 : Cancer and related data
Day2W	Num	3		Day of Birth Option 2 : Cancer and related data
Day3W	Num	3		Day of Birth Option 3 : Cancer and related data
Day4W	Num	3		Day of Birth Option 4 : Cancer and related data
Day5W	Num	3		Day of Birth Option 5 : Cancer and related data
MaoriW	Num	3	ETHM.	Ethnicity - Any Maori on Cancer, Mort, NHI, Archived NHI files
Month1W	Num	3	FMTH.	Month of Birth Option 1 : Cancer and related data
Month2W	Num	3	FMTH.	Month of Birth Option 2 : Cancer and related data
Month3W	Num	3	FMTH.	Month of Birth Option 3 : Cancer and related data
Month4W	Num	3	FMTH.	Month of Birth Option 4 : Cancer and related data
Month5W	Num	3	FMTH.	Month of Birth Option 5 : Cancer and related data
NumBthDates	Num	3		Number of Birth Dates – used for clerical review
NumMbs	Num	3		Number of Meshblocks on Cancer Data various sources – used for clerical review
NumSexonWSM	Num	3		Number of Sex codes – used for clerical review
PacFixW	Num	3	FIXP.	Pacific Ethnicity born in Pacific
PacificW	Num	3	ETHP.	Ethnicity - Any Pacific on Cancer, Mort, NHI, Archived NHI files
PossDel	Num	3		If 1 then Delete as no address and suspect not New Zealand resident
SexW	Num	3	FSEX.	Sex from WSM : Cancer and related data
Year1W	Num	4		Year of Birth Option 1 : Cancer and related data
Year2W	Num	4		Year of Birth Option 2 : Cancer and related data
Year3W	Num	4		Year of Birth Option 3 : Cancer and related data

Variable	Type	Len	Format	Label
Year4W	Num	4		Year of Birth Option 4 : Cancer and related data
Year5W	Num	4		Year of Birth Option 5 : Cancer and related data
au1w	Char	6		Area Unit Option 1 : Cancer and related data. Base 2001
au2w	Char	6		Area Unit Option 2 : Cancer and related data. Base 2001
au3w	Char	6		Area Unit Option 3 : Cancer and related data. Base 2001
au4w	Char	6		Area Unit Option 4 : Cancer and related data. Base 2001
au5w	Char	6		Area Unit Option 5 : Cancer and related data. Base 2001
au6w	Char	6		Area Unit Option 6 : Cancer and related data. Base 2001
au7w	Char	6		Area Unit Option 7 : Cancer and related data. Base 2011
au8w	Char	6		Area Unit Option 8 : Cancer and related data. Base 2001
au9w	Char	6		Area Unit Option 9 : Cancer and related data. Base 2001
cohort	Num	3		Cohort of CancerTrends – used to keep track of dataset currently being used, and used for clerical review
id_num	Char	8		CancerTrends Person Id
mb1w	Char	7		Meshblock Option 1 : Cancer and related data. Base 2001
mb2w	Char	7		Meshblock Option 1 : Cancer and related data. Base 2001
mb3w	Char	7		Meshblock Option 1 : Cancer and related data. Base 2001
mb4w	Char	7		Meshblock Option 1 : Cancer and related data. Base 2001
nonMPAW	Num	3	ETHO.	Ethnicity - Any nonMPA on Cancer, Mort, NHI, Archived NHI files
numBthGps	Num	3		Number of Countries of Birth on Cancer and related

Variable	Type	Len	Format	Label
				data – used for clerical review
numtaus	Num	3		Total Number of Area Units on Cancer Data various sources – used for clerical review

Source: (Lash 2008)

1.1.3 Notes on specific variables

Health dataset ethnicity

Ethnicity information was available on all the health datasets used in CancerTrends. Between 1981 and 2004 there was little, if any, consistency between and within these datasets about how the ethnicity information was obtained. See other publications for details on how ethnicity information was obtained over this time period for the Cancer Register and mortality collections (Shaw, Atkinson et al. 2009; Blakely, Robson et al. 2002).

1.1.3.2 Meshblock and area unit information

Mesh block (MB) and area unit (AU), also known as census area unit, are geographic units used by SNZ. A meshblock contains on average 100 people. An area unit is a bigger geographic unit, with on average 2000 people in it. Meshblocks and area units are assigned to each census form, using the stated residential address. From the point of view of the linkage process a meshblock is a better blocking variable as there are fewer people within each meshblock to match, and therefore less likelihood of a false link occurring purely by chance.

We converted all meshblocks and area units to base year 2001 for the CancerTrends linkages.

1.1.3.3 Interaction of country of birth and ethnicity

Both the census and cancer files had a country of birth and ethnicity interaction variables (AsianFix, AsianFixW, PacFix and PacFixW). These denoted Pacific and Asian people born in the Pacific or Asia respectively and was necessary to prevent the agreement weight being too high for some linkages that were not correct. (I.e. As both Pacific or Asian ethnicities are relatively uncommon in the data and Born in Pacific or Born in Asia are also relatively uncommon in the data, they both have high probability weights. If someone has both Pacific ethnicity and Born in the Pacific they will have both these weights added, and this total weight gives a disproportionate

impact on the linkage process. Therefore, we created these 'fix' variables to offset this 'double scoring'.) See previous NZCMS technical reports for more details.

1.2 Linkage Results Summary

The process of record linkage is not covered in this report. Detailed descriptions of the theory and methods of probabilistic and anonymous record linkage have been published in technical reports for the NZCMS (Hill, Atkinson et al. 2002; Fawcett, Atkinson et al. 2008). CancerTrends used largely the same process, although the software changed from the original Automatch® to the newer version called QualityStage™ for the later NZCMS linkages and all the CancerTrends linkages. The results of the CancerTrends linkage process are contained in a technical document from Statistics New Zealand (Lash 2008). Table 5 summarises the process and results of record linkage for each cohort.

Table 5 Summary of record linkage process and outcomes by cohort

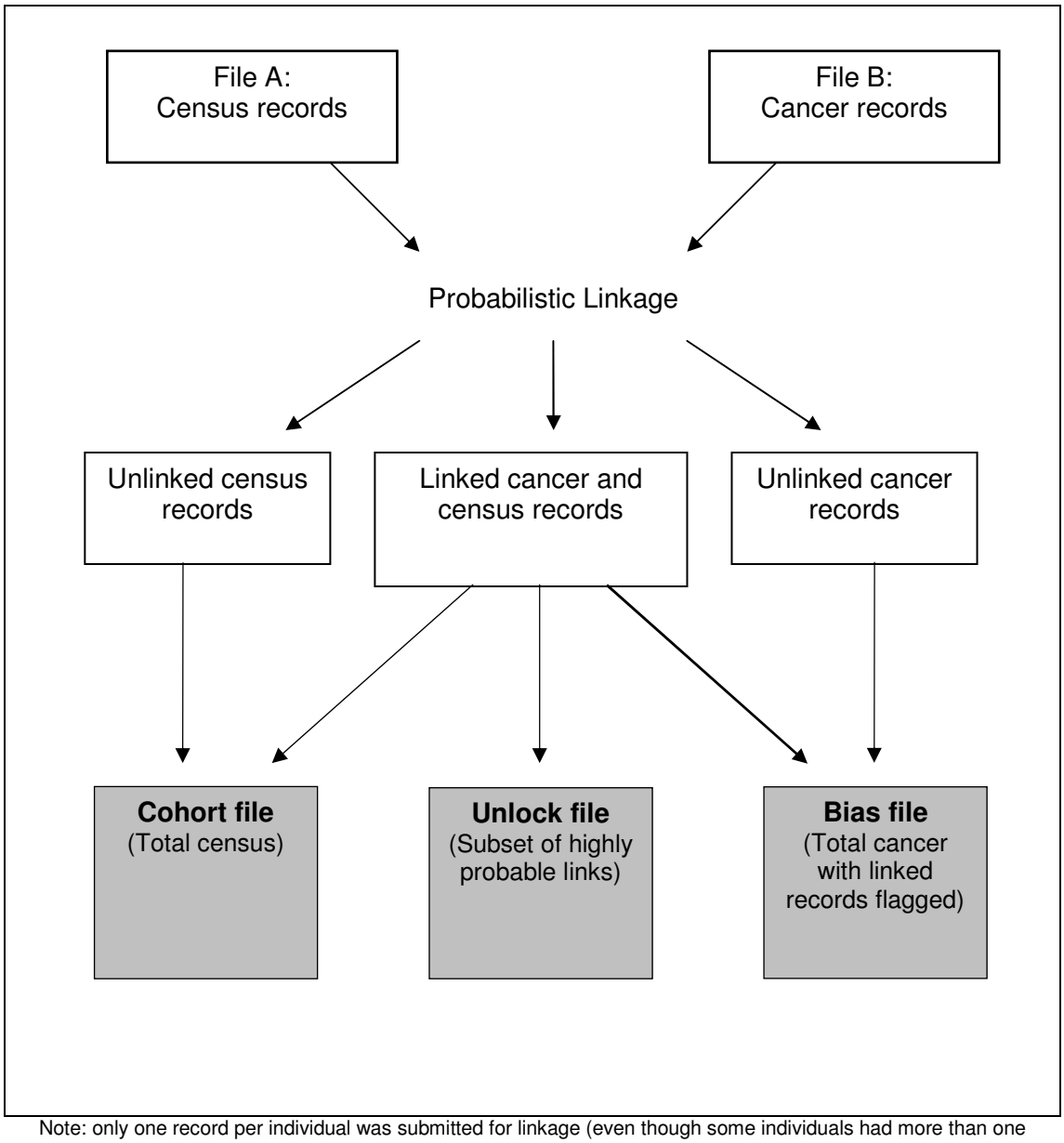
	1981-86	1986-01	1991-96	1996-01	2001-04
Usual resident population on census night (i.e. eligible to be linked)+	3 143 307	3 263 283	3 373 926	3 516 513	3 630 534
People with incident cancers from NZCR in period of follow up*	52 699	63 626	77 159	96 422	83 789
Pass 1 Links-Meshblock 1^ (Number and positive predictive value of links(%))	23780 (99.6)	30098 (99.6)	35853 (99.7)	47502 (99.6)	50312 (99.8)
Pass 2 Links-Meshblock 2^ (Number and positive predictive value of links(%))	1075 (97.4)	2339 (98.3)	2023 (98.2)	2102 (96.4)	5160 (99.2)
Pass 3 Links-Area unit 1 and sex^ (Number and positive predictive value of links (%))	5465 (84.9)	4870 (80.2)	8026 (85.3)	9058 (83.7)	4720 (75.9)
Pass 4 Links- Area unit 2 and sex^ (Number and positive predictive value of links(%))	2668 (75.8)	2956 (79.6)	5709 (78.9)	5975 (83.5)	2852 (75.4)
Linked using QualityStage™	32 988	40 263	51 611	64 637	63 044
Clerical review links	5 804	8 102	9 876	12 728	5 878
Duplicates deleted	229	244	382	474	308
Total individuals linked	38 563	49 021	61 105	76 891	68 434
Percent of people with cancer who were linked to census record	73.2%	77.1%	79.2%	79.7%	81.7%
Estimated overall positive predictive value (%) (of those linked using QualityStage™)	95.2%	95.7%	95.1%	95.8%	96.9%

^ Blocking variable . * People may have had more than one incident cancer but they were only one record per cohort was submitted for linkage. Census counts random rounded as per Statistics New Zealand policy. + Absentee records have been excluded from records available for linkage.

1.3 Final datasets

The linkage process produced three files that were released for use in the datalab at SNZ: cohort, bias and unlock datasets (see Figure 2 Summary of flow of records through linkage process).

Figure 2 Summary of flow of records through linkage process



Note: only one record per individual was submitted for linkage (even though some individuals had more than one cancer)

The **cohort** file is for future analyses of rates of cancer by social factors.
The **unlock** file is used to determine the extent of misclassification of ethnicity of cancer data compared to "gold standard" census ethnicity coding.
The **bias** file is used to calculate linkage weights to apply to the cohort file, and to correct for incomplete linkage of cancer data.

Cohort Dataset

This section looks at the specific variables in the cohort dataset which is the dataset that most analyses will be performed on and has the most detailed information from the census and the cancer register.

A description of the main variables and their associated SAS formats in the cohort dataset is contained in Appendix 1. The census derived variables are largely similar to those utilised in the NZCMS (see technical reports for further details (Fawcett, Atkinson et al. 2008; Hill, Atkinson et al. 2002)). This section contains extra information on variables from the cancer register on the cohort dataset and the process of data imputation used for some missing values on this dataset.

1.4 Cancer register variables

1.4.1 ICD Codes for cancer registration

Each cancer registration has an International Classification of Diseases (ICD) code for cancer anatomical site (ICD-10) and morphology (ICD-O). The period of time that CancerTrends covers includes transitions between different ICD coding systems for both the anatomical site and the morphological basis of disease.

1.4.1.1 Cancer anatomic site codes

All cancer diagnoses from 1980 and deaths from 1979 up until 2000 were coded by NZCR staff in ICD-9-A 2nd edition. In November 2001 NZHIS officially switched over to coding all new cancer diagnoses and deaths using ICD 10- AM 2nd edition, although many cancers from 2000 onwards were coded in this scheme (New Zealand Health Information Service 2004). See Appendix 2 for a discussion on the differences between ICD 9 and ICD 10.

All cancer registrations in the CancerTrends dataset came with both an ICD 9 and ICD 10 code for the registration, except the 80 000 registrations in the 1996 cohort which only had ICD-9 codes. This was due to the 1996 cohort data being linked a couple of years before the other cohorts to establish the feasibility of linkage. This data had to be mapped to ICD 10 to ensure consistency, the mapping process is detailed in Appendix 2.

1.4.1.2 Cancer morphology codes

Cancer morphology codes are recorded on the NZCR. Codes are obtained from the details on pathology reports, and are currently coded in the ICD-0 Version 3. Prior to 2003 they were coded in ICD-0 Version 2 (New Zealand Health Information Service 2004).

The morphology codes have been utilised as part of the classification of childhood (0-14 years) and adolescent (15-24 years) cancers (see section 1.4.1.3 for further detail)

For adults (25+ years) morphology information was retained on the datasets for brain, breast (C and D codes), testes, stomach, oesophagus, lung, trachea and bronchus. The information was retained where:

- There are a range of morphologies in that specific cancer.
- The different morphologies potentially have different aetiologies, treatment and/or prognosis.
- The cancer is common enough that dividing it by morphology will not cause confidentiality issues.

Table 6 has a summary of morphology codes over all cancers, when there are 50 or more such codes across all five cohorts. A more detailed table, and tables of morphology codes broken down by cancers, are in Appendix 4. Further work on morphology codes will be required to be able to use this information in analysis, including a determination of missingness.

Table 6 Morphology Codes present in CancerTrends data (over all cancers)

CancerTrends Morphology Codes Summary	All	Cohort of CancerTrends				
		1981	1986	1991	1996	2001
Morphology Cancer (if relevant)	281,276					
No Morphology Code		36,989	45,755	57,688	75,360	65,484
neoplasm benign/malignant	22,257	3,996	6,553	6,894	2,734	2,080
epithelial tumour benign/carcinoma in situ, NOS/carcinoma NOS	2,726	1	3	412	1,581	729
large cell carcinoma NOS	1,244	409	378	194	133	130
undifferentiated carcinoma, NOS	387	75	42	65	148	57
anaplastic carcinoma, NOS	121	66	26	18	10	1
small cell carcinoma NOS	4,176	819	840	867	888	762
oat cell carcinoma	354	131	88	57	61	17
Non-small cell carcinoma	471	.	.	.	1	470
papilloma NOS (except papilloma of bladder)/papillary carcinoma in situ/papillary carcinoma NOS	89	28	23	25	13	.

CancerTrends Morphology Codes Summary	All	Cohort of CancerTrends				
		1981	1986	1991	1996	2001
squamous cell carcinoma in situ, NOS/carcinoma NOS	10,787	2,487	2,400	2,189	2,054	1,657
squamous cell carcinoma large cell, keratinizing	532	119	77	111	150	75
squamous cell carcinoma nonkeratinizing, NOS	172	48	19	25	48	32
adenoma NOS/bronchial adenoma NOS/adenocarcinoma in situ, NOS/adenocarcinoma NOS	13,975	2,757	2,740	2,747	3,133	2,598
scirrhous adenocarcinoma	558	316	163	59	20	.
linitis plastica	84	31	13	20	8	12
adenocarcinoma intestinal type	597	2	3	49	290	253
carcinoma diffuse type	347	13	12	44	160	118
eccrine dermal cylindroma/adenoid cystic carcinoma	52	10	10	8	16	8
cribriform carcinoma	209	2	7	40	32	128
tubular adenocarcinoma	524	25	39	89	178	193
solid carcinoma, NOS	97	.	.	.	.	97
carcinoid tumour, NOS, of appendix/carcinoid tumour, NOS (except appendix m-8240)	285	46	49	40	73	77
neuroendocrine carcinoma	96	.	.	8	40	48
pulmonary adenomatosis/bronchiolo-alveolar adenocarcinoma	412	76	72	90	99	75
papillary adenoma, NOS/adencarcinoma, NOS	130	34	26	24	24	22
mucinous adenoma/mucous adenocarcinoma/pseudomyxoma peritonei	863	130	98	135	255	245
mucin-secreting adenocarcinoma	229	91	34	32	44	28
signet ring cell carcinoma/carcinoma, metastatic	454	32	56	78	148	140
intraductal carcinoma, noninfiltrating, NOS/infiltrating duct carcinoma	27,044	3,305	3,526	4,701	8,230	7,282
comedocarcinoma, noninfiltrating/NOS	286	38	26	130	67	25
intraductal papilloma/noninfiltrating adenocarcinoma, intraductal, papillary/intraductal adenocarcinoma, papillary, with invasion	269	2	11	33	108	115
intracystic adenoma, papillary/noninfiltrating carcinoma, intracystic/intracystic carcinoma, NOS	87	1	4	12	34	36
medullary carcinoma, NOS	448	103	117	84	85	59
lobular carcinoma in situ/carcinoma NOS	3,448	339	419	626	1,185	879
infiltrating ductular carcinoma	191	11	46	52	82	.
intraductal carcinoma and lobular carcinoma in situ/infiltrating duct and lobular carcinoma	837	1	.	56	309	471
Infiltrating duct mixed with other types of carcinoma	280	.	.	.	.	280
Paget"s disease mammary	108	12	19	33	31	13
Paget"s disease and infiltrating duct carcinoma of breast	257	40	36	41	83	57
Paget"s disease and intraductal carcinoma of breast	86	.	2	12	38	34
adenosquamous carcinoma	209	29	26	43	58	53
leiomyoma NOS/leiomyomatosis, NOS/leiomyosarcoma NOS	200	26	34	29	62	49
carcinosarcoma, NOS	58	2	3	4	28	21
phyllodes tumour benign/NOS/malignant	67	7	14	9	20	17
seminoma NOS	1,506	238	220	289	418	341
germinoma	50	2	8	9	17	14
embryonal carcinoma, NOS	164	18	27	33	50	36
teratoma benign/NOS/malignant, NOS	307	84	93	86	39	5

CancerTrends Morphology Codes Summary	All	Cohort of CancerTrends				
		1981	1986	1991	1996	2001
teratocarcinoma	138	34	47	35	19	3
teratoma malignant, intermediate	64	28	22	12	2	.
mixed germ cell tumour	179	.	.	5	71	103
glioma malignant	428	74	88	114	67	85
ependymoma NOS	66	11	10	7	18	20
astrocytoma NOS	1,537	458	466	388	150	75
astrocytoma anaplastic	193	8	18	24	86	57
fibrillary astrocytoma	53	5	5	11	24	8
pilocytic astrocytoma	111	10	18	17	47	19
glioblastoma NOS	1,203	32	31	179	482	479
oligodendroglioma, NOS	183	15	22	26	64	56
oligodendroglioma, anaplastic	51	1	.	4	21	25
medulloblastoma, NOS	144	33	26	26	41	18
All	384,857	53,887	65,077	79,308	99,956	86,629

1.4.1.3 Cancer groupings

Cancers were aggregated into larger, clinically relevant groups to facilitate analysis and to preserve confidentiality, especially once cancers are analysed by sex, age group, cohort and socio-economic factors. Cancer types and aetiologies vary by age, so three age appropriate classification schemes were used: children, adolescents and adults. Table 104 shows the breakdown of our final cancer groupings by age at cancer and by sex. Appendix 3 has details of the cancer groupings and details of how well they were linked by cohort and sex, plus separate tables for Children and Adolescent categories.

Adult Groupings

For adults cancers were categorised based on their site of origin. Choice of grouping was guided by other New Zealand publications on cancer (Robson, Purdie et al. 2006; New Zealand Health Information Service 2007; Ministry of Health 2002). Although there is extra detail for specific cancers, such as lung cancer and colon cancer, as numbers allowed it. As a general rule if there were less than 150 cancers per cohort in a proposed group then it was amalgamated into another group, either within the same general site or into an ‘other’ group if that was not possible. These groupings can be further amalgamated for analysis if required, but they cannot be

disaggregated. In situ cancers (ICD-10 D codes) were also kept separate where numbers allowed, to allow analysis.[2]

Children

Children (0-14) experience different cancers from adults, with some cancers having unique aetiologies and clinical prognoses. In order to facilitate understanding and reporting of childhood cancers an International Classification of Childhood Cancer (ICCC) was developed. This classification system was used for cancers in those aged 0-14 on CancerTrends in order to make the child cancer data clinically relevant, to make results internationally comparable, and to assist with SNZ confidentiality requirements.

The ICCC uses a combination of morphology and clinical site codes to group cancer in children aged 0-14 (inclusive) and has been through three iterations (Birch and Marsden 1987; Kramarova and Stiller 1996; Steliarova-Foucher, Stiller et al. 2005). The second and third editions of this classification system have been used in this project, (Kramarova and Stiller 1996; Steliarova-Foucher, Stiller et al. 2005) as morphology data on the Cancer Register changed from ICD-0 Version 2 to ICD-O Version 3 at the end of 2002 (New Zealand Health Information Service 2004). The final groupings created and the numbers in each group for each cohort are in Table 88 in Appendix 3.[3]

Adolescents

Cancer in adolescents and young people, unsurprisingly, reflects a transition between child and adult cancers, although there are also some unique features about cancers in this age group. For example the types of carcinomas that are common in 15-24 years, such as thyroid and testicular carcinoma, are uncommon in later years (Barr, Holowaty et al. 2006; Birch, Alston et al. 2002).

[2] Note CRC in situ was not included as a separate category as the numbers were too small- only 58 in the 2001 cohort. This should be reviewed if further linkages take place, due to the advent of the colorectal cancer screening programme.

[3] Cancers with very small numbers (such as retinoblastoma) had to be aggregated to preserve confidentiality. Additionally the morphology code for Langerhan's histiocytosis (histiocytosis X)- ICD-0 9722- did not have a 'home' in this classification system. There seems to be some controversy over whether this is actually cancer (Kramarova and Stiller 1996) We eventually put it in 'other' cancers of childhood, as there were only 7 cases over the entire time period.

An international classification system for cancers in 15-24 year olds is available and has been used for these age groups in CancerTrends (Barr, Holowaty et al. 2006; Birch, Alston et al. 2002; Alston, Rowan et al. 2007). The final groupings created and the numbers in each group for each cohort are in Table 89 in Appendix 3.

1.4.2 Extent of disease

Extent of disease is the Cancer Register variable that approximates clinical stage at diagnosis.[4] In 1999 there was a change in extent of disease variable, with the Cancer Register moving to the SEER Guide to Summary Staging (see Table 7) (New Zealand Health Information Service 2004). The category 'invasion of adjacent tissue/organ or regional lymph nodes' spilt into more clinically relevant 'invasion of adjacent tissue/organ' and 'regional lymph nodes'. In order to allow consistent analysis of this information in CancerTrends, SAS formats were created that combined C and D into 2 (see Table 7). The underlying values were not altered, so the SEER extent classification on the post 1 Jan 1999 registrations can be used if required.

Table 7 Change to extent of disease classification in the NZ Cancer Register

Pre 1999	**Post 1999**
0 In situ	A In situ
1 Localised and confined to organ of origin	B Localised and confined to organ of origin
2 Invasion of adjacent tissue/organ or regional lymph nodes	C Invasion of adjacent tissue/organ
	D Invasion of Regional lymph nodes

[4] There are a number of well established limitations to the extent of disease variable. Firstly there are differences in the availability of this information, with Maori being less likely to have extent recorded in many cancers, including colon, rectal, lung and breast (Robson, Purdie et al. 2006). Secondly investigation of the CancerTrends data shows that extent of disease recorded is not entirely consistent with other cancer details, for example cancers coded on ICD-10 as D codes (in-situ cancers) do not always have the 'in-situ' extent of disease code selected. Finally, extent of disease is filled out by the cancer registrars at NZHIS and is (by necessity) based on the information available to them which is pathology and laboratory specimens and death certificates, but not other investigations such as ultrasound and CT scans. A recent audit examining the accuracy of information on people with lung cancer on the NZCR showed that only 58% had the extent of disease information available (this was more likely to be missing for those with locally advanced disease, older ages or co-morbidity). For those that had the information available 77% were concordant with a hospital notes review. The discordant cases were more likely to be over staged (i.e. diagnosed with distant metastases) on the Cancer Register (Stevens, Stevens et al. 2008). An audit of colon cancer records showed a similar proportion of discrepancies between the Cancer Register and clinical records. However this review showed that the Cancer Register down staged tumours (ie they were more likely to be diagnosed with regional disease when they had metastatic) (Cunningham, Sarfati et al. 2008).

3 Distant metastases or lymph nodes	E Distant metastases or lymph nodes
5 Not known	F Not known
6 Not applicable lymphoma, leukaemia, myeloma	G Not applicable lymphoma, leukaemia, myeloma

Table 8 shows the breakdown of extent of disease classification by cancer grouping for each cohort. The classification 'Regional or node involvement' (pre1999), and the new 'Invasion of adjacent tissue or organ' and 'Regional lymph nodes' classifications have been left as separate codes for completeness but would need to be combined for any analyses over cohorts. Over all cancers there seems to be an increase over time of the percentage "not stated" and also an increase in the number "in situ" and a corresponding decrease in the number with "distant metastases", but the patterns are not consistent for each cancer grouping. A more detailed breakdown of extent of disease by cancer site is in Table 102 in Appendix 4.

Table 8 Extent of Cancer Disease by Cancer Grouping

CancerTrends Extent of Cancer Disease		All	Cohort of CancerTrends				
			1981	1986	1991	1996	2001
All	Extent of Cancer Disease						
	Not applicable (lymphomas/leukaemia/myeloma)	28,609	3,762	4,568	5,603	7,175	7,501
	Not stated	78,093	3,205	7,672	20,004	26,976	20,236
	In situ	49,227	3,611	6,482	8,296	15,636	15,202
	Localised to organ of origin	112,453	20,324	24,240	23,972	22,946	20,971
	Invasion of adjacent tissue or organ	5,910	.	.	.	2,148	3,762
	Regional lymph nodes	11,844	.	.	.	4,209	7,635
	Regional or node involvement	38,498	11,748	9,753	8,978	8,019	.
	Distant metastases	60,223	11,237	12,362	12,455	12,847	11,322
	All	384,857	53,887	65,077	79,308	99,956	86,629

Table 9 shows the breakdown of how the cancer was classified by the different cancer groupings. (Again, a more detailed breakdown by cancer site is in Appendix 4.) The vast majority are "tissue dx" (i.e. tissue diagnosis). The percentage with "Unknown" was high in the 1981 cohort, but low in later cohorts.

Table 9 Basis of Cancer Classification by Cancer Grouping

CancerTrends Basis of Cancer Classification		All	Cohort of CancerTrends				
			1981	1986	1991	1996	2001
All	Basis of Cancer Classification						
	Unknown	1,699	1,244	45	275	125	10
	Death certificate only	10,381	232	1,718	3,881	2,888	1,662
	Clinical only	8,513	1,653	2,802	2,431	1,502	125
	Investigation but no histology	17,498	2,464	2,103	3,503	4,410	5,018
	Tissue dx	346,766	48,294	58,409	69,218	91,031	79,814
	All	384,857	53,887	65,077	79,308	99,956	86,629

1.4.3 Multiple cancers

There were a total 373,715 cancers from the NZCR that were eligible to be linked and be part of the closed cohorts. A number of individuals had more than one cancer registration over the 25 year time period of the study, which could 'occur' in different cohorts or in the same cohort.

For individuals who had cancers that occurred in different cohorts no alteration was made to the records but a flag was added to their records for any cohorts after the first one to indicate that this was not their first cancer cohort (although no detail on preceding cancers is included in this flag). This will allow them to be excluded from future analysis if required.

If individuals had more than one cancer in the same cohort a few rules were applied:

- If an individual had more than one cancer registration with the **same** ICD-10 code only the first registration record for that ICD code was retained. If the subsequent cancer was registered within 4 months, the most disseminated extent of disease in either record was taken. (This is the same as NZCR extent of disease rules.)
- A very small number of people had a carcinoma in situ and an invasive carcinoma diagnosed, of the breast or cervix, within 4 months. In this case, the invasive carcinoma record was retained. (This applied only to 1 or 2 individuals.) (Note the same rule was not applied to melanoma in situ and these records were all retained.) [5]

[5] There were a number of reasons for these rules. Firstly there were concerns about whether subsequent cancers with the same ICD code were truly primary cancers, especially in the first cohorts. Secondly there were confidentiality concerns around people with large numbers of cancers. Finally it would make no difference to our analyses as incidence analyses were censored at the first cancer with the same ICD code and survival analyses will take the first cancer as it would not be clear which cancer caused the fatal event.

Table 10 summarises the number of cancers per individual before and after the application of these rules.

Table 10 Cancers per individual per cohort before and after rules applied

Number of cancers for a individual in a cohort	Number of individuals before 'rules' were applied	Number individuals after 'rules' were applied
1	360,654 (96.3%)	363,796 (97.1%)
2	13,233 (3.5%)	10,492 (2.8%)
3	656 (0.2%)	315 (0.08%)
4	58 (0.02%)	8 (0.00%)
5	4 (0.00%)	0 (0.00%)
6	4 (0.00%)	0 (0.00%)
7	1 (0.00%)	0 (0.00%)
8	1 (0.00%)	0 (0.00%)
Total	**374611 (100%)**	**374611 (100%)**

Note: These numbers include the 896 records later deleted as not NZ Residents

Table 104 in the Appendices shows the final cancer groupings we developed by age at cancer and by sex and cohort. This table can have one person represented several times in the one cohort if they developed more than one cancer. Note that because we include carcinoma in situ of the cervix (and to a lesser extent carcinoma in situ of the breast) in this and most other tables, there are many more females represented, particularly in later cohorts because of screening programs.

1.4.4 Data Imputation

Table 11 Key findings from imputation process, and recommendation to not use the imputed data.

Considerable effort was expended on imputing missing data for income and education. However, after close scrutiny of the imputed values for lung cancer it became apparent the imputation was unsatisfactory:

- the number of census respondents (regardless of cancer outcome) imputed as having high education was implausible
- the estimated rates of lung cancer among the those with imputed low income were lower than plausible, and conversely the rates of lung cancer among those with high income were higher than plausible. This pattern was more so among males, and among Māori. Consequently, rate ratios comparing lung cancer incidence among low to high income were after imputation were almost certainly biased down in many cohorts. Age- and ethnicity-standardised rates of lung cancer by income were more affected than age-only standardised, because the age- and ethnicity-standardisation 'weighted up' the spurious imputation among high income Māori.
- the estimated rate ratios comparing post-school to no qualifications tended to be much less among those with imputed data than among those with non-missing data. This apparent error was of greater magnitude among Māori.

A non-smoking related cancer, breast cancer, was similarly srcutinised. Within the limits of available data and estimation methods we used, there was no obvious difference between imputed and non-imputed data.

The possible reasons for the imputation giving rise to implausible data for lung cancer (and presumably other cancers too) include:

- not using all available data. For example, the smoking variable (whilst only available for 1981 and 1996 censuses) could have been utlised. Likewise, we should have used cancer outcome in the imputation.
- not allowing for interactions in the multiple imputation regressions. For example, given the same level of dperivation, Māori will still have the a lower income than non-Māori. By not allowing for this interaction of ethnicity with deprivation as a predictor of income, we may have biased the imputation to *overestimate* the number of Māori with high education.

- not allowing for inter-dependencies on school and tertiary qualification variables (i.e. not allowing a tertiary education variable to be imputed if the school education variable was specified as nil).

Further research on imputation in CancerTrends in warranted.

In the meantime, though, our recommendation is that cohort analyses on cancertrends data use complete data only.

The remainder of Section 1.4.4 provides a rationale for imputation, and the outputs and analyses of our attempt at multiple imputation.

1.4.4.1 A general introduction to imputation of missing data.

One issue with the census dataset is that information can be missing from the records. This is particularly problematic if data are not missing at random, and by varying degrees across the five censuses. Standard statistical methodology usually means that records missing any information on any variables will be excluded from the analysis model. This leads to the possibility of bias in estimating risk factors associated with particular variables (Vach and Blettner 2005). That is, selection bias is possible due to excluding observations with missing data. Furthermore, in situations where there are multiple variables to be included in a statistical model, the proportion of individuals excluded from the analysis can rise to levels where such selection bias becomes more likely. If the pattern of missing data differs across census cohorts, this could introduce differential bias over time – a particular problem for comparisons over time that will be a key set of cohort analyses in CancerTrends.

Multiple imputation techniques are a method for dealing with this issue by generating replacement values for missing data. Other sources of information about the individual may be able to (accurately) predict his or her characteristics on the variable that is missing data. Typically, general linear models (e.g. logistic regression) are used to calculate the likelihood of an individual falling into a certain category of the variable with missing data. This predicted value could then be substituted into the dataset, which can then be analysed using standard statistical methods.

The predictive process used will have some precision limitations, as reflected in the standard errors of the estimates. Thus the missing data substitution procedure is repeated multiple times, so that the precision of the imputation procedure is reflected in the varying values of the imputed data across these multiple datasets.

The new datasets are then analysed individually (or collectively using weights as describe later) using standard statistical methods, and the results combined so as to produce estimates of model parameters and their associated confidence intervals that take into account both the precision of the single-dataset analyses (in an identical manner to a standard statistical analysis), as well as the variability in the estimated parameters across the multiple imputed dataset analyses (reflecting how the variablility of the imputation procedure influences the estimation of the statistical models). The details of the procedure used in the CancerTrends project are outlined below. A more detailed general introduction to imputation methods is available in (Donders, van der Heijden et al. 2006) detailed theoretical and methodological considerations are presented in (Rubin 1987) and (Barnard, Rubin et al. 2005).

1.4.4.2 Creating an imputed dataset.

The aim in CancerTrends was to create imputed data to replace data originally missing with respect to ethnicity, secondary and tertiary educational attainment, and equivalised household income. These variables were selected because they are: (a) commonly used for analyses, and/or (b) had high or changing proportions of missing data over cohorts (see

Table 12). Ethnicity and educational attainment are individual level variables, whilst equivalised household income is a household level variable. Personal income was the actual variable imputed, while most cohort analyses would use equivalised household income (derived from the personal income of all household members as defined in NZCMS technical reports). Similarly, as information on secondary and tertiary educational attainment is recorded separately in the census, these variables were imputed separately, then highest qualification derived.

Table 12 shows the details of missing imputations for those aged 15 or more in the census and are usually resident in New Zealand, i.e. they are not visitors and they are not the absentee or dummy records included in the census files to account for those not recorded.

We imputed three values of personal incomes, school and tertiary education, and three ethnicity combination variables for any missing values for these people. There were not sufficient variables to conduct imputations for absentee records which only had age and sex present or already imputed by SNZ.

After doing our imputation process, we could calculate household income, but we still could not produce a household income for all people. Those with missing household incomes at the end were those with an adult absent from their usual residence on censuse night.

We calculated education separately for school and tertiary variables. Some people had one variable missing, but the other present. We should have used logical rules in addition to multiple imputation here, such that someone with no school qualifications, but missing post-school qualifications, was not able to be classified as 'post-school qualifications'.

There were not many ethnicity values that were missing. So that we only had to deal with one ethnicity variable in the imputation process, we produced a new ethnicity variable that was a combination of the main ethnicities. We had difficulty imputing using getting Proc MI in SAS, so we calculated ethnicity using a regression-based method (see later). This was a much slower process and therefore made the whole imputation process very long. In hindsight we should not have attempted to impute ethnicity – especially given the low amount of missing data that meant any added value from imputation was slight.

Table 12 shows the distribution of missing data before and after imputation.

Table 12 Percentage of adults missing data per variable by census year (Note: percentages calculated from all of the total population)

	1981-86		1986-91		1991-96		1996-01		2001-04	
Variable	N	%	N	%	N	%	N	%	N	%
Total 15+ years, Usually Resident in NZ	2,189,709	100%	2,337,675	100%	2,448,396	100%	2,588,835	100%	2,656,380	100%
Before Imputation										
Complete Income (Household and Personal) and Ethnicity data (ignoring Education variables)	1,584,429	72.4%	1,802,436	77.1%	1,883,709	76.9%	1,861,332	71.9%	1,844,418	69.4%
Complete Education (School and Tertiary) and Ethicity data (ignoring Income variables)	1,691,208	77.2%	2,231,445	95.5%	2,365,542	96.6%	2,152,629	83.2%	2,206,155	83.1%
Income										
Household Income missing before imputations	414,708	18.9%	385,113	16.5%	384,465	15.7%	475,929	18.4%	534,906	20.1%
Personal Income missing before imputations	200,097	9.1%	156,549	6.7%	195,891	8.0%	265,062	10.2%	297,423	11.2%
Education										
School Education values missing before imputations	128,505	5.9%	93,204	4.0%	76,875	3.1%	291,432	11.3%	177,645	6.7%
Tertiary Education values missing before imputations	417,213	19.1%	19,983	0.9%	60,399	2.5%	219,837	8.5%	368,364	13.9%
Either School or Tertiary Educucation values missing before imputations	484,635	22.1%	93,204	4.0%	76,878	3.1%	420,681	16.2%	436,188	16.4%
Ethnicity										
Ethnicity missing before imputations	26,967	1.2%	26,265	1.1%	20,145	0.8%	32,166	1.2%	31,662	1.2%
After Imputation										
Household Incomes still missing after imputations (unable to calculate because of adults absent on Census Night in the household and we had insufficient information to impute values for those absentees)	59,208	2.7%	152,574	6.5%	161,157	6.6%	201,099	7.8%	190,608	7.2%
Note – No Personal Income, Education or Ethnicity values missing after imputation	0		0		0		0		0	
Different Values between Imputations										
Income										
3 level Income variables have one or more different values between imputations	127,707	5.8%	80,334	3.4%	74,508	3.0%	103,542	4.0%	127,494	4.8%
5 level Income variables have one or more different values between imputations	183,687	8.4%	123,378	5.3%	114,843	4.7%	150,459	5.8%	185,310	7.0%
Education										
3 level Education variables have one or more different values between imputations	207,486	9.5%	49,152	2.1%	51,126	2.1%	176,436	6.8%	244,449	9.2%
5 level Education variables have one or more different values between imputations	233,670	10.7%	51,819	2.2%	52,944	2.2%	195,279	7.5%	266,967	10.1%
Ethnicity										
Ethnicity variables have one or more different values between imputations	6,759	0.3%	9,108	0.4%	7,356	0.3%	13,038	0.5%	12,990	0.5%

Several other variables were used as predictors in the imputation procedure: age group, sex, working status (not in workforce/part time/full time), NZDep scores, and DHB region.

Creation of imputed datasets was performed using SAS 9.1.3, through the Enterprise Guide interface running a combination of Proc MI (for those to-be-imputed variables with ordinal levels) and custom built macro routines for nominal variables (principally with respect to predicting ethnicity using a logistic regression model). Note that because the future statistical analyses that would be performed on the CancerTrends data are limited to the "Usually resident" NZ population, the imputation procedure is also limited to this group. The procedure described below was used to impute missing values for the adult population.

Step 1: For each to-be-imputed variable, missing values in the census dataset are replaced by the modal value (i.e. the most common value) calculated amongst those individuals who had intact data for that variable. For example, in the case of ethnicity, all individuals who were missing ethnicity data were initially coded as "NZ European", as this was the most common classification recorded amongst individuals reporting ethnicity.

[Steps 2 to 4 are repeated for each to-be-imputed variable]
Step 2: An appropriate statistical model is calculated for the current to-be-imputed variable (e.g. personal income). With the exception of the prediction model for ethnicity, this took the form of a cumulative logistic regression model, where the to-be-imputed variable has multiple categorical levels that can be considered ordinal (e.g. the personal income categories range from zero income to highest possible income). All other variables (as listed in the preliminary section) were used as predictor variables, with an interaction term included between age group and sex. When imputing ethnicity, the statistical model used was a multinomial logistic regression model (e.g. the multiple ethnicity categories are not ordinal), once again using all other variables as predictors with an interaction term between age group and sex.

This regression model is then applied to the current iteration of the imputed dataset for those individuals who had original data on the current to-be-imputed variable: parameters associated with each of the predictors are calculated. These predictive parameters each have a point estimate (e.g. a log odds ratio) and associated standard error term: thus information exists regarding how strongly and accurately a person's secondary educational attainment predicts the likelihood of falling into a certain personal income category.

Step 3: In plain terms, random parameter values (from a standard normal distribution) are selected from the confidence interval for each of the predictive parameters estimated in step 2. This is the first of two points in the imputation process where random elements allow the introduction of differences between the multiple imputed datasets being created. This step means that the parameter values used in the next step take into account the accuracy of the predictive model.

Step 4: For those individuals who originally had missing data on the to-be-imputed variable, a new value is created based on the parameters derived through the selection process described in Step 3. For a given individual, the likelihood of falling into each category of the to-be-imputed variable can be calculated using the parameters calculated in Steps 2 and 3, and the current values of the predictor variables in the imputation model for that individual (e.g. the likelihood of falling into each possible income bracket can be defined given that we have this individual's values on age group, sex, working status, NZDep, ethnicity, educational status, and DHB region). A random number is drawn (from a uniform distribution between 0 and 1) which allows the selection of the new value for the to-be-imputed variable. Note that this is the second point in the imputation procedure where differences between multiple imputed datasets are created.
[Steps 2 to 4 are now repeated for the other to-be-imputed variables]

Once all variables have been imputed once, the imputation process has finished one iteration. The originally missing variables have all been replaced with imputed values based on the predictive models outlined above. Some of the variables (e.g. working status, if this was imputed first) will have been imputed at a point in the process where the modal value is still being used for the predictive side of the equation for all the other to-be-imputed variables; the predictive model used for the last variable imputed in this iteration (e.g. ethnicity) has been imputed with predictive values for all other to-be-imputed variables having been imputed once already. Therefore the iterative process is repeated several times to allow the estimates to stabilise. For this project, five iterations were used to allow this stabilisation to occur.

At the end of this process, imputation has been completed for a single dataset: (Rubin 1987) gives a formula for estimating the number of imputed datasets that are required to give accurate coverage of the true confidence intervals for the main

analyses. Given a composite missing data rate of 20% for the CancerTrends census data (see

Table 12 for proportions of individuals missing data by census year), it was calculated that 3 datasets should be used for this analysis.

1.4.4.3 Intended cohort analyses using the imputed datasets.

The analysis procedure for imputed data proceeds in two stages. In stage one, data analysis proceeds as per standard analysis (e.g. Poisson regression), and these analyses are repeated on each dataset. The parameter estimates (e.g. log rate ratios for a Poisson regression) and associated standard errors are retained from each analysis. In stage two, the results from the three imputed datasets are then combined on a parameter-by-parameter basis to calculate a point estimate of the parameter as well as an associated confidence interval. The formulae for the steps described below are standard, as reported by (Rubin 1987) p. 76; see also SAS help documents, for Proc MIANALYZE (SAS Institute 2000-2004). When estimating parameters where the statistical modelling depends on a log or logit transformation, such as a rate ratio or odds ratio, these stage two calculations are performed on the transformed parameters through to deriving the confidence interval, at which point the log or logit values are transformed back to rate ratios/odds ratios.

For the results from Stage one, the parameter estimates in each imputed dataset will differ due to the random processes in the dataset creation procedure. The aim of stage two of the data analysis procedure is to use the variability between the imputed datasets to produce feasible estimates of the true confidence interval around each parameter. The following steps are therefore repeated separately for each parameter in the current analysis model. The results from the three individually performed analyses at stage one will thus provide three different versions of the parameter estimate, and three different versions of the variance associated with this parameter estimate (this variance would usually be used to construct a confidence interval around the parameter estimate if this was a single dataset analysis).

The final combined value of the parameter estimate is simply calculated by taking the mean value of the three parameter estimates across the separate analyses from stage one. The average variance associated with the parameter across the three datasets is also calculated, giving the mean within-imputation variance. The difference between each of the three individual parameter estimates and the final combined parameter estimate is also calculated, and used to produce a measure of

between-imputation variance across the three analyses: this is therefore a measure of how much the parameter estimates in each of the three imputed datasets differ from one another.

These two sources of variance are then used to produce a total variance associated with this particular parameter estimate: the mean parameter estimate and total variance are then used to derive a final confidence interval for the parameter (formulae presented in (Rubin 1987) p. 76-77, and repeated in (Barnard, Rubin et al. 2005), with the width of this confidence interval being dependent on a t-distribution with a set number of degrees of freedom (dependent on the number of imputed datasets used for the analysis).

For rate ratios estimated outside the context of a regression model, the combination of rate ratios across imputed datasets are calculated in a similar manner: rate ratios and the associated variance are calculated separately for each imputed dataset, and then the across-imputation combination (as described above for regression methods) procedure is applied to this information to provide a rate ratio and confidence interval that includes information regarding how variable the estimates were across the multiple imputation sets.

Running all standardised rate ratio analyses three times, one per each imputation dataset, for a range of cancers, socio-economic variables, age groupings, etc, on CancerTrends data in the Datalab is extremely time consuming. Therefore, we investigated using Fractionally Weighted Imputation (FWI) (Fay 1996) as a quicker alternative, using the original imputation values as produced by the Rubin multiple imputation method. With FWI all imputed and non-imputed data is analysed at the same time, with the imputation values being assigned a weight of 1/*m*, which in our case with m=3 datasets was 1/3. This weight was multiplied by our linkage bias weight (*w_agethadj*) to produce a combined weight (*w_agethfwi*).

If we calculated the FWI rates without any adjustment of the variance, we found that the confidence intervals were slightly too narrow and needed to be adjusted. (This was due to having three records for imputation values implying greater certainty, whereas they should only be treated as one record.) Therefore, we multipled the standardised variance by 1/(proportion unimputed events), i.e.

FWI Variance Adjustment =

$$\frac{\text{total events (non-weighted)}}{}$$

(total events (non-weighted) – number events imputed (non-weighted))

The resultant Fractionally Weighted Imputation rates, rate ratios and rate differences and associated confidence intervals were very similar to the Rubin imputation method, but were produced substantially quicker, and will therefore be the preferred method used for future cohort analyses.

1.4.4.4 Scrutiny of imputed datasets and results.

The following tables (Table 13 - Table 17) show a comparison between restricting the data to those records without missing values (i.e. complete records), imputation 1, 2, 3 and summary using Rubin's imputation method, and FWI method for Lung Cancer, Males, 25+ years, equivalised household income 3 levels. (Table 18 - Table 22 show corresponding comparisons for highest qualification 3 levels.)

And among the imputed data analyses, there was negligible difference between the "imputed summary" and "FWI" results, supporting the future use of the FWI method.

However, there is a major problem with the imputed data – especially for Māori – and as forewarned in Table 11 above. The tables include a row labelled "Est. Imputed Only". The standardiused rate (SR) in these rows is that back-estimated from the change in the SR from that for complete data to that for complete plus missing data. For example, in the first instance of this in Table 13 below, the SR estimate of 114 is given by {(118 x 1,276,106) - (120 x 906,694)} / (1,276,106 – 906,694). Examining these SR estimates in Table 13 to Table 22 (lung cancer, males) and in Appendix 7 (lung cancer, females) the patterns described in Table 11 on page 45 are apparent. More directly, if one examines and compares the SRR among the the "Complete Income and Eth" and "Est Imputed Only" large incosistencies are noted, and large enough on many occasions to substantially (and implausibly in our view) lower the SRR in the "Imputation Summary" row.

Table 13 Imputation Method comparisions of standardised rate (SR), standardised rate ratio (SRR) and standardised rate differences (SRD) by equivalised household income for Lung Cancer, Males, Income, 1981-86

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males 25+ years*		*1981-86*				
Income	*By: All Ethnicities*		*Age Standardised*				
Low Income	Complete Income and Eth.	1575	.	906,694	120 (112 - 127)	1.41 (1.28 - 1.56)	35 (25 - 45)
	Est. Imputed Only	339	17.7	369,412	114 (-4.9% diff)	1.39 (-1.4% diff)	32 (-9.1% diff)
	Imputation 1	1902	17.2	1,275,724	118 (111 - 124)	1.41 (1.29 - 1.53)	34 (26 - 42)
	Imputation 2	1920	18.0	1,276,290	118 (112 - 125)	1.40 (1.29 - 1.52)	34 (25 - 42)
	Imputation 3	1917	17.8	1,276,303	118 (112 - 125)	1.41 (1.29 - 1.53)	34 (26 - 43)
	Imputation Summary	1914	17.7	1,276,106	118 (112 - 124)	1.41 (1.29 - 1.53)	34 (26 - 42)
	FWI Analysis	1914	17.7	1,276,106	118 (112 - 124)	1.40 (1.30 - 1.52)	34 (26 - 42)
Medium Income	Complete Income and Eth.	936	.	1,083,977	90.7 (83.8 - 97.6)	1.07 (0.96 - 1.19)	6.1 (-3.4 - 16)
	Est. Imputed Only	399	29.9	411,288	95.1 (4.8% diff)	1.16 (8.5% diff)	13 (115.5% diff)
	Imputation 1	1350	30.7	1,495,326	92.9 (87.1 - 98.8)	1.11 (1.01 - 1.22)	9.2 (1.2 - 17)
	Imputation 2	1317	28.9	1,494,956	91.1 (85.2 - 96.9)	1.08 (0.98 - 1.18)	6.5 (-1.5 - 15)
	Imputation 3	1335	29.9	1,495,511	91.8 (85.9 - 97.6)	1.09 (1.00 - 1.20)	7.7 (-0.3 - 16)
	Imputation Summary	1335	29.9	1,495,265	91.9 (85.7 - 98.1)	1.09 (0.99 - 1.20)	7.8 (-0.8 - 16)
	FWI Analysis	1335	29.9	1,495,265	91.9 (86.1 - 97.7)	1.09 (1.00 - 1.19)	7.8 (0.0 - 16)
High Income	Complete Income and Eth.	1053	.	1,156,025	84.6 (78.2 - 91.1)	1	0
	Est. Imputed Only	300	22.2	264,530	81.9 (-3.2% diff)	1	0
	Imputation 1	1344	21.7	1,420,875	83.7 (78.2 - 89.2)	1	0
	Imputation 2	1365	22.9	1,420,679	84.6 (79.1 - 90.1)	1	0
	Imputation 3	1350	22.0	1,420,111	84.0 (78.5 - 89.5)	1	0
	Imputation Summary	1353	22.2	1,420,555	84.1 (78.5 - 89.7)	1	0
	FWI Analysis	1353	22.2	1,420,555	84.1 (78.9 - 89.4)	1	0
Income	*By: All Ethnicities*		*Age and Ethnicity Standardised*				
Low Income	Complete Income and Eth.	1575	.	906,694	126 (116 - 136)	1.41 (1.23 - 1.62)	37 (23 - 51)
	Est. Imputed Only	339	17.7	369,412	109 (-13.4% diff)	1.18 (-16.3% diff)	17 (-54.8% diff)
	Imputation 1	1902	17.2	1,275,724	121 (113 - 129)	1.35 (1.20 - 1.51)	31 (20 - 43)
	Imputation 2	1920	18.0	1,276,290	121 (113 - 129)	1.33 (1.19 - 1.49)	30 (19 - 42)
	Imputation 3	1917	17.8	1,276,303	121 (113 - 129)	1.36 (1.22 - 1.52)	32 (21 - 44)
	Imputation Summary	1914	17.7	1,276,106	121 (113 - 129)	1.35 (1.20 - 1.51)	31 (20 - 43)
	FWI Analysis	1914	17.7	1,276,106	121 (113 - 129)	1.35 (1.21 - 1.50)	31 (20 - 42)
Medium Income	Complete Income and Eth.	933	.	1,083,977	93.5 (83.7 - 103)	1.05 (0.90 - 1.22)	4.3 (-9.6 - 18)
	Est. Imputed Only	399	30.0	411,288	105 (12.1% diff)	1.13 (8.0% diff)	12 (187.2% diff)
	Imputation 1	1350	30.9	1,495,326	97.9 (89.7 - 106)	1.10 (0.97 - 1.24)	8.6 (-3.1 - 20)
	Imputation 2	1314	29.0	1,494,956	95.3 (87.1 - 103)	1.05 (0.92 - 1.19)	4.1 (-7.6 - 16)
	Imputation 3	1332	30.0	1,495,511	96.5 (88.4 - 105)	1.08 (0.96 - 1.23)	7.4 (-4.1 - 19)
	Imputation Summary	1332	30.0	1,495,265	96.6 (87.9 - 105)	1.08 (0.94 - 1.23)	6.7 (-6.1 - 20)
	FWI Analysis	1332	30.0	1,495,265	96.5 (88.7 - 104)	1.07 (0.96 - 1.21)	6.7 (-4.3 - 18)
High Income	Complete Income and Eth.	1053	.	1,156,025	89.2 (79.3 - 99.1)	1	0
	Est. Imputed Only	300	22.2	264,530	92.4 (3.6% diff)	1	0
	Imputation 1	1347	21.8	1,420,875	89.3 (81.0 - 97.6)	1	0
	Imputation 2	1365	22.9	1,420,679	91.1 (82.7 - 99.5)	1	0
	Imputation 3	1350	22.0	1,420,111	89.1 (81.0 - 97.2)	1	0
	Imputation Summary	1353	22.2	1,420,555	89.8 (81.2 - 98.5)	1	0
	FWI Analysis	1353	22.2	1,420,555	89.8 (82.2 - 97.5)	1	0

Table 14 Imputation Method comparisons for Lung Cancer, Males, Income, 1986-91

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1986-91*			
Income	*By: All Ethnicities*			*Age Standardised*			
Low Income	Complete Income and Eth.	1638	.	1,117,132	107 (101 - 113)	1.59 (1.44 - 1.74)	40 (32 - 48)
	Est. Imputed Only	156	8.7	307,897	105 (-1.7% diff)	1.34 (-15.6% diff)	27 (-32.4% diff)
	Imputation 1	1785	8.2	1,424,974	106 (101 - 112)	1.54 (1.41 - 1.68)	37 (30 - 45)
	Imputation 2	1803	9.2	1,424,867	108 (102 - 113)	1.56 (1.43 - 1.71)	39 (31 - 46)
	Imputation 3	1791	8.5	1,425,245	107 (101 - 112)	1.54 (1.41 - 1.69)	38 (30 - 45)
	Imputation Summary	1794	8.7	1,425,028	107 (101 - 113)	1.55 (1.41 - 1.69)	38 (30 - 46)
	FWI Analysis	1791	8.5	1,425,072	107 (101 - 112)	1.55 (1.42 - 1.69)	38 (30 - 45)
Medium Income	Complete Income and Eth.	1377	.	1,302,720	92.5 (86.9 - 98.0)	1.37 (1.24 - 1.51)	25 (17 - 33)
	Est. Imputed Only	270	16.4	355,609	102 (10.1% diff)	1.30 (-5.3% diff)	23 (-6.2% diff)
	Imputation 1	1656	16.8	1,658,819	95.0 (89.8 - 100)	1.38 (1.26 - 1.51)	26 (19 - 33)
	Imputation 2	1641	16.1	1,658,479	94.0 (88.9 - 99.2)	1.37 (1.25 - 1.50)	25 (18 - 32)
	Imputation 3	1650	16.5	1,657,689	94.6 (89.4 - 99.8)	1.37 (1.25 - 1.50)	26 (19 - 33)
	Imputation Summary	1647	16.4	1,658,329	94.5 (89.2 - 99.8)	1.37 (1.26 - 1.50)	26 (19 - 33)
	FWI Analysis	1650	16.5	1,658,320	94.6 (89.3 - 99.9)	1.37 (1.25 - 1.51)	26 (18 - 33)
High Income	Complete Income and Eth.	879	.	1,158,754	67.6 (62.4 - 72.8)	1	0
	Est. Imputed Only	174	16.5	157,461	78.5 (16.1% diff)	1	0
	Imputation 1	1050	16.3	1,315,780	68.9 (64.0 - 73.7)	1	0
	Imputation 2	1050	16.3	1,316,227	68.7 (63.9 - 73.6)	1	0
	Imputation 3	1056	16.8	1,316,639	69.0 (64.2 - 73.9)	1	0
	Imputation Summary	1053	16.5	1,316,215	68.9 (64.0 - 73.7)	1	0
	FWI Analysis	1050	16.3	1,316,181	68.9 (63.9 - 73.9)	1	0
Income	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
Low Income	Complete Income and Eth.	1638	.	1,117,132	114 (105 - 123)	1.57 (1.36 - 1.81)	41 (29 - 53)
	Est. Imputed Only	153	8.5	307,897	99.4 (-12.6% diff)	0.91 (-42.1% diff)	-9.9 (-124.1% diff)
	Imputation 1	1788	8.4	1,424,974	110 (103 - 117)	1.42 (1.24 - 1.61)	32 (21 - 44)
	Imputation 2	1800	9.0	1,424,867	111 (104 - 119)	1.47 (1.29 - 1.67)	36 (25 - 47)
	Imputation 3	1788	8.4	1,425,245	111 (103 - 118)	1.43 (1.26 - 1.63)	33 (22 - 45)
	Imputation Summary	1791	8.5	1,425,028	111 (103 - 118)	1.44 (1.26 - 1.65)	34 (22 - 46)
	FWI Analysis	1794	8.7	1,425,072	111 (104 - 118)	1.44 (1.27 - 1.63)	34 (23 - 44)
Medium Income	Complete Income and Eth.	1374	.	1,302,720	103 (94 - 112)	1.43 (1.23 - 1.65)	31 (18 - 43)
	Est. Imputed Only	273	16.6	355,609	114 (10.4% diff)	1.04 (-27.0% diff)	4.8 (-84.3% diff)
	Imputation 1	1656	17.0	1,658,819	106 (98 - 114)	1.36 (1.19 - 1.57)	28 (16 - 40)
	Imputation 2	1641	16.3	1,658,479	105 (97 - 114)	1.39 (1.22 - 1.59)	30 (18 - 41)
	Imputation 3	1647	16.6	1,657,689	106 (97 - 114)	1.37 (1.19 - 1.57)	29 (16 - 41)
	Imputation Summary	1647	16.6	1,658,329	106 (97 - 114)	1.37 (1.20 - 1.58)	29 (17 - 41)
	FWI Analysis	1650	16.7	1,658,320	106 (98 - 114)	1.37 (1.21 - 1.57)	29 (17 - 40)
High Income	Complete Income and Eth.	879	.	1,158,754	72.5 (64.0 - 81.1)	1	0
	Est. Imputed Only	174	16.5	157,461	109 (50.7% diff)	1	0
	Imputation 1	1050	16.3	1,315,780	77.7 (68.9 - 86.4)	1	0
	Imputation 2	1050	16.3	1,316,227	75.8 (67.7 - 84.0)	1	0
	Imputation 3	1056	16.8	1,316,639	77.2 (68.7 - 85.8)	1	0
	Imputation Summary	1053	16.5	1,316,215	76.9 (68.1 - 85.7)	1	0
	FWI Analysis	1050	16.3	1,316,181	76.9 (68.8 - 85.0)	1	0

Table 15 Imputation Method comparisons for Lung Cancer, Males, Income, 1991-96

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1991-96*			
Income	*By: All Ethnicities*			*Age Standardised*			
Low Income	Complete Income and Eth.	1125	.	1,073,870	105 (98 - 112)	1.61 (1.46 - 1.78)	40 (31 - 48)
	Est. Imputed Only	114	9.2	264,651	101 (-2.9% diff)	1.62 (0.6% diff)	39 (-2.0% diff)
	Imputation 1	1242	9.4	1,338,351	104 (97 - 111)	1.62 (1.47 - 1.78)	40 (32 - 48)
	Imputation 2	1236	9.0	1,338,388	104 (97 - 110)	1.61 (1.46 - 1.77)	39 (31 - 47)
	Imputation 3	1242	9.4	1,338,823	104 (97 - 111)	1.60 (1.46 - 1.76)	39 (31 - 47)
	Imputation Summary	1239	9.2	1,338,520	104 (97 - 111)	1.61 (1.46 - 1.77)	39 (31 - 47)
	FWI Analysis	1239	9.2	1,338,507	104 (97 - 111)	1.61 (1.46 - 1.77)	39 (31 - 48)
Medium Income	Complete Income and Eth.	2109	.	1,364,068	91.0 (86.3 - 95.8)	1.40 (1.28 - 1.54)	26 (19 - 33)
	Est. Imputed Only	174	7.6	379,883	88.7 (-2.5% diff)	1.42 (1.1% diff)	26 (-0.2% diff)
	Imputation 1	2283	7.6	1,744,025	90.6 (86.3 - 95.0)	1.41 (1.30 - 1.53)	26 (20 - 33)
	Imputation 2	2286	7.7	1,743,924	90.8 (86.5 - 95.2)	1.41 (1.30 - 1.53)	26 (20 - 33)
	Imputation 3	2277	7.4	1,743,903	90.0 (85.7 - 94.3)	1.39 (1.28 - 1.51)	25 (19 - 31)
	Imputation Summary	2283	7.6	1,743,951	90.5 (86.0 - 94.9)	1.40 (1.29 - 1.53)	26 (20 - 32)
	FWI Analysis	2280	7.5	1,743,907	90.5 (86.4 - 94.7)	1.40 (1.29 - 1.53)	26 (20 - 32)
High Income	Complete Income and Eth.	864	.	1,289,456	64.9 (60.0 - 69.8)	1	0
	Est. Imputed Only	198	18.6	280,366	62.7 (-3.5% diff)	1	0
	Imputation 1	1059	18.4	1,569,918	64.3 (59.9 - 68.6)	1	0
	Imputation 2	1065	18.9	1,569,981	64.5 (60.2 - 68.9)	1	0
	Imputation 3	1068	19.1	1,569,567	64.8 (60.5 - 69.2)	1	0
	Imputation Summary	1062	18.6	1,569,822	64.5 (60.1 - 68.9)	1	0
	FWI Analysis	1062	18.6	1,569,801	64.5 (59.8 - 69.2)	1	0
Income	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
Low Income	Complete Income and Eth.	1128	.	1,073,870	109 (101 - 117)	1.41 (1.22 - 1.64)	32 (19 - 45)
	Est. Imputed Only	111	9.0	264,651	100 (-7.9% diff)	1.14 (-19.0% diff)	12 (-60.8% diff)
	Imputation 1	1242	9.2	1,338,351	107 (100 - 115)	1.39 (1.22 - 1.59)	30 (19 - 42)
	Imputation 2	1236	8.7	1,338,388	107 (99 - 114)	1.34 (1.17 - 1.53)	27 (15 - 39)
	Imputation 3	1239	9.0	1,338,823	107 (100 - 115)	1.34 (1.17 - 1.53)	27 (15 - 39)
	Imputation Summary	1239	9.0	1,338,520	107 (100 - 114)	1.36 (1.18 - 1.56)	28 (16 - 41)
	FWI Analysis	1239	9.0	1,338,507	107 (100 - 114)	1.36 (1.18 - 1.56)	28 (16 - 40)
Medium Income	Complete Income and Eth.	2109	.	1,364,068	98.8 (92.1 - 106)	1.28 (1.11 - 1.49)	22 (9.7 - 34)
	Est. Imputed Only	171	7.5	379,883	98.3 (-0.5% diff)	1.12 (-12.3% diff)	11 (-51.1% diff)
	Imputation 1	2280	7.5	1,744,025	99.0 (92.9 - 105)	1.28 (1.13 - 1.46)	22 (11 - 33)
	Imputation 2	2283	7.6	1,743,924	99.3 (93.2 - 106)	1.25 (1.09 - 1.42)	20 (8.7 - 31)
	Imputation 3	2274	7.3	1,743,903	97.7 (91.6 - 104)	1.22 (1.07 - 1.39)	18 (6.6 - 29)
	Imputation Summary	2280	7.5	1,743,951	98.7 (92.2 - 105)	1.25 (1.08 - 1.44)	20 (7.8 - 32)
	FWI Analysis	2280	7.5	1,743,907	98.7 (92.9 - 104)	1.25 (1.09 - 1.43)	20 (8.7 - 31)
High Income	Complete Income and Eth.	867	.	1,289,456	77.0 (66.9 - 87.1)	1	0
	Est. Imputed Only	195	18.4	280,366	87.6 (13.8% diff)	1	0
	Imputation 1	1059	18.1	1,569,918	77.1 (68.2 - 86.0)	1	0
	Imputation 2	1062	18.4	1,569,981	79.5 (70.3 - 88.8)	1	0
	Imputation 3	1065	18.6	1,569,567	80.0 (70.7 - 89.3)	1	0
	Imputation Summary	1062	18.4	1,569,822	78.9 (69.1 - 88.6)	1	0
	FWI Analysis	1065	18.6	1,569,801	78.9 (69.3 - 88.4)	1	0

Table 16 Imputation Method comparisons for Lung Cancer, Males, Income, 1996-01

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1996-01*			
Income	*By: All Ethnicities*			*Age Standardised*			
Low Income	Complete Income and Eth.	1473	.	1,162,046	97.1 (91.5 - 103)	1.88 (1.70 - 2.08)	45 (38 - 52)
	Est. Imputed Only	189	11.4	289,926	101 (4.1% diff)	1.73 (-8.1% diff)	43 (-6.1% diff)
	Imputation 1	1665	11.5	1,451,694	98.1 (92.8 - 103)	1.81 (1.66 - 1.98)	44 (38 - 51)
	Imputation 2	1653	10.9	1,452,213	97.4 (92.2 - 103)	1.82 (1.67 - 1.99)	44 (38 - 51)
	Imputation 3	1665	11.5	1,452,010	98.1 (92.8 - 103)	1.84 (1.68 - 2.01)	45 (38 - 51)
	Imputation Summary	1662	11.4	1,451,972	97.9 (92.5 - 103)	1.82 (1.67 - 2.00)	44 (38 - 51)
	FWI Analysis	1662	11.4	1,451,972	97.9 (93.4 - 102)	1.82 (1.68 - 1.98)	44 (38 - 50)
Medium Income	Complete Income and Eth.	1572	.	1,220,907	82.7 (77.7 - 87.6)	1.60 (1.45 - 1.77)	31 (25 - 38)
	Est. Imputed Only	246	13.5	451,413	80.5 (-2.7% diff)	1.38 (-14.0% diff)	22 (-29.1% diff)
	Imputation 1	1803	12.8	1,672,705	81.5 (77.2 - 85.9)	1.51 (1.38 - 1.65)	28 (22 - 33)
	Imputation 2	1830	14.1	1,672,137	82.8 (78.4 - 87.2)	1.55 (1.42 - 1.69)	29 (24 - 35)
	Imputation 3	1818	13.5	1,672,119	82.1 (77.8 - 86.5)	1.54 (1.41 - 1.68)	29 (23 - 34)
	Imputation Summary	1818	13.5	1,672,320	82.1 (77.5 - 86.7)	1.53 (1.40 - 1.68)	29 (22 - 35)
	FWI Analysis	1815	13.4	1,672,304	82.1 (78.4 - 85.9)	1.53 (1.41 - 1.66)	29 (23 - 34)
High Income	Complete Income and Eth.	699	.	1,308,417	51.7 (47.5 - 55.9)	1	0
	Est. Imputed Only	264	27.4	507,579	58.5 (13.1% diff)	1	0
	Imputation 1	969	27.9	1,815,889	54.1 (50.3 - 57.9)	1	0
	Imputation 2	957	27.0	1,815,939	53.4 (49.6 - 57.1)	1	0
	Imputation 3	960	27.2	1,816,159	53.4 (49.7 - 57.2)	1	0
	Imputation Summary	963	27.4	1,815,996	53.6 (49.8 - 57.5)	1	0
	FWI Analysis	963	27.4	1,815,999	53.6 (49.9 - 57.3)	1	0
Income	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
Low Income	Complete Income and Eth.	1473	.	1,162,046	105 (98 - 113)	1.93 (1.70 - 2.19)	51 (41 - 60)
	Est. Imputed Only	189	11.4	289,926	97.8 (-7.1% diff)	1.46 (-24.1% diff)	31 (-38.7% diff)
	Imputation 1	1668	11.7	1,451,694	104 (98 - 110)	1.74 (1.56 - 1.94)	44 (36 - 53)
	Imputation 2	1653	10.9	1,452,213	103 (97 - 110)	1.79 (1.61 - 2.00)	46 (38 - 54)
	Imputation 3	1665	11.5	1,452,010	104 (98 - 111)	1.83 (1.64 - 2.04)	47 (39 - 56)
	Imputation Summary	1662	11.4	1,451,972	104 (97 - 110)	1.79 (1.58 - 2.02)	46 (37 - 55)
	FWI Analysis	1662	11.4	1,451,972	104 (98 - 110)	1.79 (1.62 - 1.98)	46 (38 - 53)
Medium Income	Complete Income and Eth.	1572	.	1,220,907	92.4 (85.6 - 99.3)	1.69 (1.49 - 1.92)	38 (29 - 47)
	Est. Imputed Only	246	13.5	451,413	84.6 (-8.4% diff)	1.27 (-25.0% diff)	18 (-52.8% diff)
	Imputation 1	1806	13.0	1,672,705	89.1 (83.4 - 94.8)	1.49 (1.34 - 1.67)	30 (22 - 37)
	Imputation 2	1830	14.1	1,672,137	91.4 (85.6 - 97.2)	1.59 (1.42 - 1.77)	34 (26 - 42)
	Imputation 3	1818	13.5	1,672,119	90.3 (84.6 - 96.0)	1.58 (1.42 - 1.76)	33 (26 - 41)
	Imputation Summary	1818	13.5	1,672,320	90.3 (84.0 - 96.6)	1.55 (1.36 - 1.78)	32 (23 - 42)
	FWI Analysis	1815	13.4	1,672,304	90.2 (85.2 - 95.3)	1.55 (1.41 - 1.72)	32 (25 - 39)
High Income	Complete Income and Eth.	696	.	1,308,417	54.6 (48.9 - 60.4)	1	0
	Est. Imputed Only	267	27.7	507,579	66.8 (22.3% diff)	1	0
	Imputation 1	969	28.2	1,815,889	59.6 (54.3 - 64.9)	1	0
	Imputation 2	960	27.5	1,815,939	57.5 (52.4 - 62.7)	1	0
	Imputation 3	960	27.5	1,816,159	57.0 (52.0 - 62.1)	1	0
	Imputation Summary	963	27.7	1,815,996	58.0 (52.0 - 64.1)	1	0
	FWI Analysis	963	27.7	1,815,999	58.0 (53.2 - 62.9)	1	0

Table 17 Imputation Method comparisons for Lung Cancer, Males, Income, 2001-04

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*2001-04*			
Income	*By: All Ethnicities*			*Age Standardised*			
Low Income	Complete Income and Eth.	1167	.	840,524	93.9 (87.8 - 100)	1.95 (1.75 - 2.18)	46 (38 - 53)
	Est. Imputed Only	183	13.6	258,764	91.8 (-2.3% diff)	1.59 (-18.7% diff)	34 (-26.0% diff)
	Imputation 1	1356	13.9	1,099,472	94.1 (88.5 - 99.7)	1.86 (1.69 - 2.04)	43 (37 - 50)
	Imputation 2	1350	13.6	1,099,515	93.2 (87.6 - 98.8)	1.85 (1.68 - 2.03)	43 (36 - 49)
	Imputation 3	1344	13.2	1,098,877	93.0 (87.4 - 98.6)	1.83 (1.67 - 2.02)	42 (36 - 49)
	Imputation Summary	1350	13.6	1,099,288	93.4 (87.7 - 99.2)	1.85 (1.68 - 2.04)	43 (36 - 50)
	FWI Analysis	1350	13.6	1,099,277	93.4 (88.6 - 98.3)	1.84 (1.68 - 2.02)	43 (37 - 49)
Medium Income	Complete Income and Eth.	1047	.	830,978	71.2 (66.2 - 76.3)	1.48 (1.32 - 1.66)	23 (17 - 30)
	Est. Imputed Only	201	16.1	375,430	72.2 (1.4% diff)	1.25 (-15.7% diff)	14 (-38.4% diff)
	Imputation 1	1239	15.5	1,205,674	71.0 (66.6 - 75.5)	1.40 (1.27 - 1.54)	20 (15 - 26)
	Imputation 2	1254	16.5	1,206,460	71.8 (67.3 - 76.3)	1.42 (1.29 - 1.57)	21 (16 - 27)
	Imputation 3	1254	16.5	1,207,087	71.6 (67.1 - 76.0)	1.41 (1.28 - 1.55)	21 (15 - 27)
	Imputation Summary	1248	16.1	1,206,407	71.5 (66.9 - 76.0)	1.41 (1.28 - 1.56)	21 (15 - 27)
	FWI Analysis	1251	16.3	1,206,410	71.5 (67.5 - 75.4)	1.41 (1.28 - 1.55)	21 (15 - 26)
High Income	Complete Income and Eth.	573	.	1,150,973	48.1 (43.8 - 52.4)	1	0
	Est. Imputed Only	279	32.7	417,225	57.9 (20.3% diff)	1	0
	Imputation 1	852	32.7	1,568,747	50.7 (46.9 - 54.4)	1	0
	Imputation 2	852	32.7	1,567,918	50.5 (46.7 - 54.2)	1	0
	Imputation 3	855	33.0	1,567,929	50.8 (47.0 - 54.5)	1	0
	Imputation Summary	852	32.7	1,568,198	50.7 (46.9 - 54.4)	1	0
	FWI Analysis	852	32.7	1,568,181	50.6 (46.8 - 54.4)	1	0
Income	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
Low Income	Complete Income and Eth.	1170	.	840,524	98.9 (91.5 - 106)	1.81 (1.57 - 2.08)	44 (34 - 54)
	Est. Imputed Only	180	13.3	258,764	92.5 (-6.4% diff)	1.48 (-18.3% diff)	30 (-32.3% diff)
	Imputation 1	1359	13.9	1,099,472	98.4 (91.9 - 105)	1.76 (1.57 - 1.96)	42 (34 - 51)
	Imputation 2	1350	13.3	1,099,515	97.3 (90.9 - 104)	1.70 (1.52 - 1.91)	40 (32 - 48)
	Imputation 3	1344	12.9	1,098,877	96.4 (90.1 - 103)	1.69 (1.51 - 1.89)	39 (31 - 48)
	Imputation Summary	1350	13.3	1,099,288	97.4 (90.6 - 104)	1.72 (1.52 - 1.94)	41 (32 - 50)
	FWI Analysis	1350	13.3	1,099,277	97.3 (91.9 - 103)	1.71 (1.53 - 1.92)	41 (33 - 48)
Medium Income	Complete Income and Eth.	1047	.	830,978	81.9 (74.3 - 89.5)	1.50 (1.28 - 1.74)	27 (17 - 37)
	Est. Imputed Only	204	16.3	375,430	73.5 (-10.2% diff)	1.17 (-21.7% diff)	11 (-59.7% diff)
	Imputation 1	1242	15.7	1,205,674	79.0 (73.0 - 85.0)	1.41 (1.25 - 1.59)	23 (15 - 31)
	Imputation 2	1254	16.5	1,206,460	78.9 (73.0 - 84.8)	1.38 (1.22 - 1.55)	22 (14 - 30)
	Imputation 3	1254	16.5	1,207,087	80.0 (73.9 - 86.0)	1.40 (1.24 - 1.58)	23 (15 - 31)
	Imputation Summary	1251	16.3	1,206,407	79.3 (73.1 - 85.5)	1.40 (1.24 - 1.58)	23 (14 - 31)
	FWI Analysis	1248	16.1	1,206,410	79.3 (73.9 - 84.6)	1.40 (1.24 - 1.57)	23 (15 - 30)
High Income	Complete Income and Eth.	573	.	1,150,973	54.7 (48.0 - 61.4)	1	0
	Est. Imputed Only	279	32.7	417,225	62.6 (14.4% diff)	1	0
	Imputation 1	852	32.7	1,568,747	56.0 (50.9 - 61.1)	1	0
	Imputation 2	849	32.5	1,567,918	57.2 (51.8 - 62.6)	1	0
	Imputation 3	855	33.0	1,567,929	57.1 (51.7 - 62.5)	1	0
	Imputation Summary	852	32.7	1,568,198	56.8 (51.3 - 62.3)	1	0
	FWI Analysis	852	32.7	1,568,181	56.8 (51.3 - 62.3)	1	0

Table 18 Imputation Method comparisons of standardised rate (SR), standardised rate ratio (SRR) and standardised rate differences (SRD) by highest qualification for Lung Cancer, Males, Education, 1981-86

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1981-86*			
Education	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
No Qualifications	Complete Educ. and Eth.	2856	.	1,845,706	112 (106 - 118)	1.64 (1.29 - 2.09)	44 (27 - 61)
	Est. Imputed Only	768	21.2	469,647	101 (-10.1% diff)	1.13 (-31.0% diff)	12 (-73.1% diff)
	Imputation 1	3615	21.0	2,315,423	110 (104 - 115)	1.54 (1.29 - 1.83)	38 (25 - 51)
	Imputation 2	3633	21.4	2,315,720	110 (105 - 115)	1.47 (1.19 - 1.81)	35 (19 - 51)
	Imputation 3	3624	21.2	2,314,916	110 (105 - 115)	1.48 (1.22 - 1.79)	36 (21 - 50)
	Imputation Summary	3624	21.2	2,315,353	110 (105 - 115)	1.50 (1.23 - 1.83)	36 (21 - 51)
	FWI Analysis	3624	21.2	2,315,353	110 (105 - 115)	1.50 (1.21 - 1.86)	37 (20 - 53)
School Qualifications	Complete Educ. and Eth.	291	.	419,693	107 (79 - 135)	1.57 (1.10 - 2.22)	39 (6.5 - 71)
	Est. Imputed Only	87	23.0	125,351	58.7 (-45.1% diff)	0.66 (-57.9% diff)	-30 (-178.0% diff)
	Imputation 1	381	23.6	545,022	101 (79 - 123)	1.42 (1.08 - 1.86)	30 (5.1 - 54)
	Imputation 2	375	22.4	543,938	92.6 (74.0 - 111)	1.24 (0.93 - 1.65)	18 (-6.4 - 42)
	Imputation 3	378	23.0	546,170	94.1 (75.5 - 113)	1.27 (0.97 - 1.66)	20 (-3.1 - 43)
	Imputation Summary	378	23.0	545,043	95.9 (73.8 - 118)	1.31 (0.95 - 1.80)	23 (-5.6 - 51)
	FWI Analysis	378	23.0	545,043	96.1 (73.2 - 119)	1.31 (0.95 - 1.81)	23 (-4.7 - 51)
Post-School Qualifications	Complete Educ. and Eth.	456	.	1,084,297	68.3 (52.2 - 84.4)	1	0
	Est. Imputed Only	234	33.9	355,861	88.9 (30.2% diff)	1	0
	Imputation 1	693	34.2	1,440,109	71.2 (59.4 - 83.1)	1	0
	Imputation 2	684	33.3	1,440,896	74.9 (59.4 - 90.3)	1	0
	Imputation 3	690	33.9	1,439,469	74.2 (60.6 - 87.8)	1	0
	Imputation Summary	690	33.9	1,440,158	73.4 (59.0 - 87.8)	1	0
	FWI Analysis	687	33.6	1,440,158	73.2 (57.7 - 88.6)	1	0

Table 19 Imputation Method comparisons for Lung Cancer, Males, Education, 1986-91

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1986-91*			
Education	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
No Qualifications	Complete Educ. and Eth.	2418	.	1,746,194	107 (101 - 112)	1.39 (1.21 - 1.61)	30 (19 - 42)
	Est. Imputed Only	252	9.4	104,338	122 (15.0% diff)	0.54 (-61.5% diff)	-106 (-452.9% diff)
	Imputation 1	2685	9.9	1,850,825	108 (102 - 113)	1.33 (1.15 - 1.55)	27 (14 - 40)
	Imputation 2	2664	9.2	1,850,429	107 (102 - 113)	1.33 (1.14 - 1.54)	26 (13 - 39)
	Imputation 3	2664	9.2	1,850,341	107 (102 - 113)	1.32 (1.13 - 1.54)	26 (13 - 39)
	Imputation Summary	2670	9.4	1,850,532	107 (102 - 113)	1.33 (1.14 - 1.55)	26 (14 - 39)
	FWI Analysis	2670	9.4	1,850,532	107 (103 - 112)	1.33 (1.19 - 1.48)	27 (17 - 36)
School Qualifications	Complete Educ. and Eth.	795	.	897,743	100 (83.2 - 117)	1.31 (1.06 - 1.62)	24 (4.0 - 43)
	Est. Imputed Only	66	7.7	35,918	45.4 (-54.6% diff)	0.20 (-84.8% diff)	-183 (-873.4% diff)
	Imputation 1	846	6.0	933,395	97.8 (82.9 - 113)	1.21 (0.98 - 1.49)	17 (-1.7 - 36)
	Imputation 2	870	8.6	933,868	98.4 (84.0 - 113)	1.22 (0.99 - 1.49)	18 (-1.0 - 36)
	Imputation 3	864	8.0	933,720	97.6 (82.4 - 113)	1.20 (0.97 - 1.49)	16 (-2.8 - 36)
	Imputation Summary	861	7.7	933,661	97.9 (83.1 - 113)	1.21 (0.98 - 1.49)	17 (-1.9 - 36)

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
	FWI Analysis	861	7.7	933,661	97.9 (83.0 - 113)	1.21 (1.01 - 1.45)	17 (0.0 - 34)
Post-School Qualifications	Complete Educ. and Eth.	1080	.	1,814,594	76.4 (66.2 - 86.6)	1	0
	Est. Imputed Only	96	8.2	55,249	229 (199.3% diff)	1	0
	Imputation 1	1176	8.2	1,869,815	80.7 (69.3 - 92.2)	1	0
	Imputation 2	1173	7.9	1,869,738	80.9 (69.2 - 92.6)	1	0
	Imputation 3	1179	8.4	1,869,975	81.2 (69.5 - 93.0)	1	0
	Imputation Summary	1176	8.2	1,869,843	80.9 (69.3 - 92.6)	1	0
	FWI Analysis	1176	8.2	1,869,843	80.9 (72.6 - 89.2)	1	0

Table 20 Imputation Method comparisons for Lung Cancer, Males, Education, 1991-96

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1991-96*			
Education	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
No Qualifications	Complete Educ. and Eth.	2259	.	1,592,198	110 (105 - 116)	1.32 (1.17 - 1.49)	27 (16 - 37)
	Est. Imputed Only	90	3.8	51,547	87.8 (-20.3% diff)	1.67 (26.3% diff)	35 (31.5% diff)
	Imputation 1	2352	4.0	1,644,104	109 (104 - 115)	1.33 (1.19 - 1.49)	27 (17 - 37)
	Imputation 2	2355	4.1	1,643,661	110 (104 - 115)	1.36 (1.21 - 1.52)	29 (19 - 39)
	Imputation 3	2337	3.3	1,643,469	109 (104 - 114)	1.30 (1.16 - 1.45)	25 (15 - 35)
	Imputation Summary	2349	3.8	1,643,745	109 (104 - 115)	1.33 (1.18 - 1.50)	27 (16 - 38)
	FWI Analysis	2349	3.8	1,643,745	109 (105 - 114)	1.33 (1.20 - 1.47)	27 (18 - 36)
School Qualifications	Complete Educ. and Eth.	1080	.	1,014,510	82.7 (74.8 - 90.7)	0.99 (0.86 - 1.15)	-0.7 (-13 - 12)
	Est. Imputed Only	30	2.7	26,043	46.7 (-43.5% diff)	0.89 (-10.4% diff)	-5.9 (746.7% diff)
	Imputation 1	1107	2.4	1,040,075	82.3 (74.1 - 90.5)	1	0
	Imputation 2	1113	3.0	1,040,358	81.4 (73.8 - 89.1)	1.01 (0.88 - 1.16)	0.6 (-11 - 12)
	Imputation 3	1107	2.4	1,041,227	81.8 (74.0 - 89.5)	0.97 (0.85 - 1.12)	-2.2 (-14 - 9.3)
	Imputation Summary	1110	2.7	1,040,553	81.8 (73.9 - 89.7)	0.99 (0.86 - 1.15)	-0.5 (-13 - 11)
	FWI Analysis	1110	2.7	1,040,553	81.8 (74.6 - 88.9)	0.99 (0.87 - 1.13)	-0.6 (-11 - 9.9)
Post-School Qualifications	Complete Educ. and Eth.	1308	.	2,172,475	83.4 (74.2 - 92.6)	1	0
	Est. Imputed Only	96	6.8	73,066	52.7 (-36.9% diff)	1	0
	Imputation 1	1401	6.6	2,245,660	82.2 (74.0 - 90.5)	1	0
	Imputation 2	1395	6.2	2,245,820	80.9 (72.8 - 89.0)	1	0
	Imputation 3	1419	7.8	2,245,144	84.0 (75.4 - 92.6)	1	0
	Imputation Summary	1404	6.8	2,245,541	82.4 (73.3 - 91.4)	1	0
	FWI Analysis	1404	6.8	2,245,541	82.4 (74.7 - 90.0)	1	0

Table 21 Imputation Method comparisons for Lung Cancer, Males, Education, 1996-01

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*1996-01*			
Education	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
No Qualifications	Complete Educ. and Eth.	2346	.	1,647,261	98.5 (93.4 - 104)	1.76 (1.48 - 2.11)	43 (32 - 53)
	Est. Imputed Only	432	15.6	271,919	92.9 (-5.7% diff)	1.08 (-38.7% diff)	6.7 (-84.2% diff)
	Imputation 1	2793	16.0	1,919,798	98.2 (93.6 - 103)	1.49 (1.33 - 1.68)	33 (24 - 41)

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
	Imputation 2	2781	15.6	1,919,153	97.8 (93.2 - 102)	1.54 (1.38 - 1.72)	34 (27 - 42)
	Imputation 3	2763	15.1	1,918,590	97.1 (92.5 - 102)	1.51 (1.35 - 1.69)	33 (25 - 41)
	Imputation Summary	2778	15.6	1,919,180	97.7 (92.9 - 103)	1.51 (1.34 - 1.71)	33 (25 - 42)
	FWI Analysis	2778	15.6	1,919,180	97.7 (93.5 - 102)	1.51 (1.33 - 1.72)	33 (24 - 42)
School Qualifications	Complete Educ. and Eth.	735	.	1,274,648	69.1 (62.1 - 76.1)	1.24 (1.02 - 1.51)	13 (1.5 - 25)
	Est. Imputed Only	228	23.7	219,229	86.1 (24.7% diff)	1.00 (-19.3% diff)	0.0 (-99.9% diff)
	Imputation 1	951	22.7	1,493,213	69.7 (63.8 - 75.6)	1.06 (0.92 - 1.22)	4.0 (-5.3 - 13)
	Imputation 2	966	23.9	1,494,083	71.7 (65.5 - 77.9)	1.13 (0.99 - 1.29)	8.3 (-0.6 - 17)
	Imputation 3	975	24.6	1,494,334	73.4 (67.1 - 79.7)	1.14 (1.00 - 1.31)	9.2 (0.1 - 18)
	Imputation Summary	963	23.7	1,493,877	71.6 (64.2 - 79.0)	1.11 (0.94 - 1.30)	7.2 (-3.9 - 18)
	FWI Analysis	963	23.7	1,493,877	71.6 (65.8 - 77.5)	1.11 (0.96 - 1.28)	7.1 (-2.7 - 17)
Post-School Qualifications	Complete Educ. and Eth.	450	.	1,343,223	55.8 (46.3 - 65.3)	1	0
	Est. Imputed Only	486	51.9	540,454	86.1 (54.3% diff)	1	0
	Imputation 1	936	51.9	1,883,723	65.7 (58.5 - 72.8)	1	0
	Imputation 2	930	51.6	1,883,499	63.5 (57.1 - 69.8)	1	0
	Imputation 3	939	52.1	1,883,811	64.3 (57.7 - 70.8)	1	0
	Imputation Summary	936	51.9	1,883,677	64.5 (57.3 - 71.7)	1	0
	FWI Analysis	936	51.9	1,883,677	64.5 (56.7 - 72.3)	1	0

Table 22 Imputation Method comparisons for Lung Cancer, Males, Education, 2001-04

Variable	Imputation Method	Number Cancers	%Imp	Person Time (yrs)	SR (95% CI)	SRR (95% CI)	SRD (95% CI)
Lung	*Males*	*25+ years*		*2001-04*			
Education	*By: All Ethnicities*			*Age and Ethnicity Standardised*			
No Qualifications	Complete Educ. and Eth.	1482	.	1,034,890	93.4 (87.6 - 99.3)	1.96 (1.66 - 2.31)	46 (36 - 55)
	Est. Imputed Only	507	25.5	267,781	76.4 (-18.2% diff)	0.81 (-58.9% diff)	-18 (-140.0% diff)
	Imputation 1	2016	26.5	1,302,246	90.9 (86.1 - 95.7)	1.61 (1.40 - 1.85)	34 (26 - 43)
	Imputation 2	1977	25.0	1,302,466	89.4 (84.6 - 94.1)	1.74 (1.55 - 1.96)	38 (31 - 45)
	Imputation 3	1977	25.0	1,303,301	89.5 (84.8 - 94.3)	1.64 (1.46 - 1.85)	35 (27 - 43)
	Imputation Summary	1989	25.5	1,302,671	89.9 (84.8 - 95.1)	1.66 (1.42 - 1.94)	36 (27 - 45)
	FWI Analysis	1989	25.5	1,302,671	89.9 (85.3 - 94.6)	1.67 (1.48 - 1.88)	36 (29 - 43)
School Qualifications	Complete Educ. and Eth.	621	.	932,383	75.9 (66.7 - 85.1)	1.59 (1.31 - 1.94)	28 (17 - 40)
	Est. Imputed Only	234	27.4	168,675	96.8 (27.5% diff)	1.02 (-35.7% diff)	2.1 (-92.6% diff)
	Imputation 1	846	26.6	1,100,896	76.3 (68.8 - 83.7)	1.35 (1.15 - 1.58)	20 (9.5 - 30)
	Imputation 2	867	28.4	1,101,806	81.9 (73.3 - 90.5)	1.60 (1.38 - 1.85)	31 (21 - 41)
	Imputation 3	852	27.1	1,100,472	79.1 (70.8 - 87.4)	1.45 (1.25 - 1.68)	25 (14 - 35)
	Imputation Summary	855	27.4	1,101,058	79.1 (68.8 - 89.4)	1.46 (1.14 - 1.87)	25 (9.1 - 41)
	FWI Analysis	855	27.4	1,101,058	79.0 (71.4 - 86.7)	1.46 (1.27 - 1.69)	25 (15 - 35)
Post-School Qualifications	Complete Educ. and Eth.	555	.	1,489,079	47.7 (40.4 - 54.9)	1	0
	Est. Imputed Only	267	32.5	234,761	94.7 (98.5% diff)	1	0
	Imputation 1	804	31.0	1,724,427	56.5 (49.4 - 63.7)	1	0
	Imputation 2	825	32.7	1,723,297	51.2 (46.0 - 56.5)	1	0
	Imputation 3	837	33.7	1,723,796	54.6 (48.7 - 60.4)	1	0
	Imputation Summary	822	32.5	1,723,840	54.1 (45.5 - 62.7)	1	0
	FWI Analysis	825	32.7	1,723,840	54.0 (48.2 - 59.8)	1	0

Unlock dataset and calculation of unlock ratios

This section details the production of the calculation of ratios (known as unlock ratios) to correct for misclassification of ethnicity on the New Zealand Cancer register, namely undercounting of Māori, Pacific and Asian peoples and over counting of nMPA. The first section details the unlock dataset, then the methods to calculate the ratios are described and finally the results. Further detail on the theory and previous methods used in the calculation of unlock ratios for mortality data on the NZCMS is available elsewhere (Ajwani, Blakely et al. 2003; Fawcett, Atkinson et al. 2008). A summary paper of key results for the cancer registry-census discrepancies for 1981 to 2004 has been published elsewhere (Shaw, Atkinson et al. 2009), and will be presented in more detail here.

1.5 Unlock dataset

A detailed description of the variables in the unlock dataset is in Appendix 4. In summary the variables include demographic variables plus ethnicity values from both the census and the health datasets for the same person. This enables us to compare ethnicity on census and cancer files for what should be the same person. See NZCMS publications for full details of restrictions to generate the highly probable links (HPL) needed for calculating unlock ratios.

1.6 Methods to calculate unlock ratios

1.6.1 Classifying ethnicity

Ethnicity would be simple to classify in the absence of multiple responses. However ethnic identity in New Zealand has become increasingly more diverse, with substantial numbers of people identifying with more than one ethnic group (at least on the census) (Callister and Blakely 2004).

A number of different methods for ethnicity output are available to approach multiple ethnic group affiliations. These are sole ethnicity, prioritised ethnicity, single/combination and total ethnicity (see Table 23).[6] The rationale and uses for each of these approaches is different, and more detailed discussion is available

[6] Note these are different methods of outputting the same information.

elsewhere (Robson and Reid 2001; Statistics New Zealand 2004; Statistics New Zealand 2005; Statistics New Zealand 2005; Callister and Blakely 2004; Blakely, Tobias et al. 2007).

Table 23 Different ethnicity output approaches

Approach	*Definition*
Sole	*Those who only identify with one ethnic group are placed in that ethnic group eg sole Māori, sole Pacific, sole Asian or sole NZ European. Those who report more than one affiliation are excluded from analysis using this approach.*
Prioritised	*Those who affiliate with more than one ethnicity are placed in only one ethnic group in a priority order of Māori, Pacific, Asian and then other. Each individual is only in one group.*
Single/Combination output	*This places people either into a single ethnic group if they only had one response (i.e. sole ethnic groups) or into the appropriate combination group if that is what they indicated. E.g. someone who said they were Pacific, Māori and NZ European would be in a combination group of Pacific/Māori/NZ European. Each individual is only in one group.*
Total	*Total ethnicity places an individual in all ethnic groups that they identify with, thus capturing multiple ethnic affiliations of individuals. If an individual indicates any/all of Maori, Pacific, Asian and/or NZ European ethnic affiliation they were placed in any/all of Total Māori, Total Pacific, Total Asian ethnic groups. Thus the sum of the ethnic groups is greater than the number of people. Note that if people identify with more than one Pacific ethnic group – ie Samoan and Tongan-or Asian ethnic group they will only be recorded in Pacific or Asian once respectively.*

Source: (Robson and Reid 2001; Statistics New Zealand 2005; Callister and Blakely 2004)

Unlock ratios have been calculated for prioritised and (a modified version of) total ethnicity. The modification of total ethnicity in that people who did not indicate any of Māori, Pacific or Asian ethnic affiliations were placed in the non-Māori/Pacific/Asian (nMPA). The latter group is not, strictly speaking, a 'total' ethnic group, as in a true total approach individuals who indicated that they were, for example, affiliated both with Māori and NZ European/pakeha ethnic groups should be recorded in both groups. However in order to have a reference group that did not overlap with any/all other ethnic groups (necessary for the calculation of standard errors and the ability to

make meaningful comparisons with a reference group) we made this a residual category.

The 1981 ethnicity question was a blood quantum question. We separated the fractions into three component ethnicity variables and then treated these variables the same as all the other cohorts.

1.6.2 Calculating the extent of misclassification of ethnicity on the cancer register

The methods described below for calculating unlock ratios are a modification of, and improvement over, previous methods used in the NZCMS to be able to cope with 5 large cohorts of data (Fawcett, Atkinson et al. 2008).

Highly probable links (HPL), those census-cancer record links that were exact matches without ethnicity in the matching process (61.6%, 67.6%, 71.2%,75.7% and 69.5% for each consecutive cohort), were used to calculate unlock ratios. The observations on the HPL dataset were then weighted up to be representative of the total eligible cancer registration population. This weighting required specifying the 'best' stratification of the datasets by socio-demographic characteristics (sex, age, ethnicity, territorial local authority, NZDep, rurality and time since census) to capture variability in the likelihood of being in the HPL dataset. We used iterative regression modelling to select these strata aiming to identify strata that identified substantial variation in the probability of being in the HPL, and conversely and most importantly inform the aggregation of adjacent strata (necessary due to small numbers) only when there was no meaningful or likely variation in the probability of being an HPL.

The weighting method was very similar to that used for the Bias linkage weights (see 1.9). The SAS programmes in Appendix 12 to Appendix 14 give a complete record of the process of creating the unlock weights and calculating the unlock ratios.

- We tried to have strata levels as detailed as possible but needed to combine some due to small numbers.
- Sex was always treated as separate strata, it was never combined.
- Aggregation was necessary to maintain numbers – every strata needed some HPL (unlock) records, it could not just have non-HPL records.

- We used people with cancers for the strata and weights, not separate cancer records (people could have up to 4 cancers per cohort). They were linked independently of what cancers they had.
- Our principle was to try and have consistent strata and regimes across all 5 unlock datasets with only minor differences between groupings, but we also wanted the strata to be as detailed as possible.
- We used regression analyses separately for each of the 5 bias datasets using the unlockflag i.e. if they were in the HPL or Unlock dataset as the outcome variable. We did some initial regressions (not reported – see authors for details) to identify strata for collapsing – and finally decided on the following regressions. Results of these regressions are in Table 106 and Table 107.
    - Regression was logistic using Proc Genmod in SAS.
    - We save the results of the Type I and Type III tests and made one dataset from all 5 regressions.
    - Saved all parameter estimates into one dataset.
    - Looked at patterns over all analyses to determine which variables and levels were important or which ones could be combined.
    - Variables in regression were
        - cenyear (always the same within each dataset), e.g. 1981, 1986, 1991, 1996, 2001)
        - sex (male=1, female=2)
        - Age (approximately 10 year age groups: ' 0-14 yrs', '15-24 yrs', '25-29 yrs', '30-34 yrs', '35-39 yrs', '40-49 yrs', '50-59 yrs', '60-69 yrs', '70-79 yrs', '>=80 years')
        - Maori on C1 (Maori ethnicity on first cancer diagnosis, not Maori and missing ethnicity)
        - Pacific on C1 (Pacific ethnicity on first cancer diagnosis, not Pacific and missing ethnicity)
        - Asian on C1 (Asian ethnicity on first cancer diagnosis, not Asian and missing ethnicity)
        - nonMPA on C1 (non-Maori non-Pacific non-Asian ethnicity (also called European/Other) on first cancer diagnosis, not relevant and missing ethnicity)
        - Time since census (Number of months after census before first cancer (made negative so that the date closest to census is the reference group) 0='Dates -ve, 0- 6 mths', -7=' 7-12 mths', -

13='13-18 mths', -19='19-24 mths', -25='25-30 mths', -31='31-36 mths', -37='37-42 mths', -43='43-high mths')

- Territorial authority (Invercargill made reference group. 3 cohorts also had unknown territorial authority field)
- NZ Deprivation index (3 groups Deciles 1–6 (ref), Deciles 7-8, Deciles 9-10). Had to use NZDep 2001 as the record linkage used the base 2001 versions of meshblock and area unit.
- Area Unit mobility (proportion of area unit that is mobile, grouped into 4 groups Decile 1-4 (ref), Decile 5-6, Decile 7-8, Decile 9-10, and missing).
- Dependent variable was unlockflag (0=not on unlock dataset, 1=has highly probable link therefore in unlock dataset)

- To save time with too much manual groupings of strata that did not meet our criteria, for the final production we had a preferred level for each variable and an alternative less detailed level that could be used if necessary. These differed slightly for the different cohorts.
  - Final groupings for TA were done on statistical significance to the reference group. (It has now been pointed out that these groupings should not have been just on statistical significance but size of parameter estimates should have been used as well.)
  - Used results of regression to get an idea of relative importance of variables as well as ways the variables could be amalgamated.
  - Final order of variables in decending order of importance was sex, age, TA, ethnicity, time since census, rurality, NZDep.
  - Needed to put ethnicity into prioritised ethnicity to give it just one value for each person.
  - For strata groupings over the 5 cohorts :
    - Sex was always kept separate.
    - Age was always kept at level 1.
    - Territorial Authority was kept at grouping level 1 for 1981, grouping level 1 with alternative level 2 for 1986, 1991, 1996, and grouping level 1 with alternative level 4 for 2001.
    - Prioritised ethnicity was level 1 for 1981 and 1991, grouping level 1 with alternative all grouped together for 1996 and 2001, and grouping level 3 with alternative all grouped together for 1986.

- Time since census was level 1 with alternative all grouped together for 1981 and 1991, level 3 with alternative all for 1996 and 2001, and all grouped together for 1986.
- Rurality was level 4 with alternative all grouped together for 1991, and for all other cohorts all levels were grouped together.
- NZ Deprivation was all grouped together for all cohorts.

The SAS programme for this data management and analysis above can be found in Appendix 12 (CreateUnlockWgtonBiasFinal.sas).

The final stratification of each HPL data set aimed to achieve as many strata as possible to capture variation in the proportion of registrants in the HPL data set compared to all people eligible (556, 200, 1235, 258, 341 individual strata for each of the five consecutive cohorts), yet ensuring that all strata had at least one HPL record. Median records in each strata for each cohort consecutively were 12 (maximum 1978), 37.5 (maximum 3080), 13 (maximum 983), 16 (maximum 5563), and 35 (maximum 2883). Inverse probability weights were then assigned to each strata, e.g. if there were 20 eligible male cancer registrations aged 45-64 in a specific strata and 15 of these were in the HPL dataset, then each of these 15 was given a weight of 1.33 (i.e. 20/15).

Table 24 Main Strata groupings for HPL Unlock Weights

Dataset	**Variable and level**	**Levels**
1981	Age Level 1	0-14 yrs, 15-24 yrs, 25-29 yrs, 30-34 yrs, 35-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 1	Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 +ve Prob, 0.10 –ve Prob, 0.05 +ve Prob, 0.05 –ve Prob, 0.01 +ve Prob, 0.01 –ve Prob, Missing
	Prioritised Ethnicity Level 1	Maori, Pacific, Asian, NonMaoriNonPacificNonMissing, Missing
	Time since census Level 1 (with alternative All Combined)	*First Option* : Dates –ve&0-6 mths, 7-12 mths, 13-18 mths, 19-24 mths, 25-30 mths, 31-36 mths, 37-42 mths, 43-high

		mths, Missing *Alternative* : All Combined
	Rurality All Combined	All Combined
	NZ Dep All Combined	All Combined
1986	Age Level 1	0-14 yrs, 15-24 yrs, 25-29 yrs, 30-34 yrs, 35-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 1 (with alternative Level 2)	*First Option* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 +ve Prob, 0.10 –ve Prob, 0.05 +ve Prob, 0.05 –ve Prob, 0.01 +ve Prob, 0.01 –ve Prob, Missing *Alternative* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 Prob, 0.05 Prob, 0.01 Prob, Missing
	Prioritised Ethnicity Level 3 (with alternative All Combined)	*First Option* : Maori, Pacific, NonMaoriNonPacific *Alternative* : All Combined
	Time since census All Combined	All Combined
	Rurality All Combined	All Combined
	NZ Dep All Combined	All Combined
1991	Age Level 1	0-14 yrs, 15-24 yrs, 25-29 yrs, 30-34 yrs, 35-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 1 (with alternative Level 2)	*First Option* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 +ve Prob, 0.10 –ve Prob, 0.05 +ve Prob, 0.05 –ve Prob, 0.01 +ve Prob, 0.01 –ve Prob, Missing *Alternative* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 Prob, 0.05 Prob, 0.01 Prob, Missing
	Prioritised Ethnicity Level 1	Maori, Pacific, Asian, NonMaoriNonPacificNonMissing, Missing

	Time since census Level 1 (with alternative All Combined)	*First Option* : Dates –ve&0-6 mths, 7-12 mths, 13-18 mths, 19-24 mths, 25-30 mths, 31-36 mths, 37-42 mths, 43-high mths, Missing *Alternative* : All Combined
	Rurality Level 4 (with alternative All Combined)	All Urban, NonUrban or Missing *Alternative* : All Combined
	NZ Dep All Combined	All Combined
1996	Age Level 1	0-14 yrs, 15-24 yrs, 25-29 yrs, 30-34 yrs, 35-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 1 (with alternative Level 2)	*First Option* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 +ve Prob, 0.10 –ve Prob, 0.05 +ve Prob, 0.05 –ve Prob, 0.01 +ve Prob, 0.01 –ve Prob, Missing *Alternative* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 Prob, 0.05 Prob, 0.01 Prob, Missing
	Prioritised Ethnicity Level 1 (with alternative All Combined)	Maori, Pacific, Asian, NonMaoriNonPacificNonMissing, Missing *Alternative* : All Combined
	Time since census Level 3 (with alternative All Combined)	*First Option* : Dates –ve&0-24 mths, 25-high mths *Alternative* : All Combined
	Rurality All Combined	All Combined
	NZ Dep All Combined	All Combined
2001	Age Level 1	0-14 yrs, 15-24 yrs, 25-29 yrs, 30-34 yrs, 35-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 1 (with alternative Level 4)	*First Option* : Statistical Significance : 0.99 Prob, 0.70 Prob, 0.50 Prob, 0.10 +ve Prob, 0.10 –ve Prob, 0.05 +ve Prob, 0.05 –ve Prob, 0.01 +ve Prob,

		0.01 –ve Prob, Missing *Alternative* : Statistical Significance : N.S.&0.10, 0.01 or 0.05 Prob, Missing
	Prioritised Ethnicity Level 1 (with alternative All Combined)	Maori, Pacific, Asian, NonMaoriNonPacificNonMissing, Missing *Alternative* : All Combined
	Time since census Level 3 (with alternative All Combined)	*First Option* : Dates –ve&0-24 mths, 25-high mths *Alternative* : All Combined
	Rurality All Combined	All Combined
	NZ Dep All Combined	All Combined

Once the datasets were weighted up to be representative of the cancer registrant population we then cross classified the number of cancer registrants by their ethnic group codes on both cancer and census data.

1.7 Results

1.7.1 Ratios

Results for total ethnicity are presented here. Interpretation can be found in the NZMJ paper (Shaw, Atkinson et al. 2009). Prioritised ethnicity tables are available in Appendix 4.

Table 25 Cross classified cancers, misclassification ratios for all cancers by total ethnicity and cohort 1981-2004

Ethnicity	1981 Census ethnicity	Cancer Register ethnicity	Ratio^	1986 Census ethnicity	Cancer Register ethnicity	Ratio^	1991 Census ethnicity	Cancer Register ethnicity	Ratio^	1996 Census ethnicity	Cancer Register ethnicity	Ratio^	2001 Census ethnicity	Cancer Register ethnicity	Ratio^
Total Māori	2,829	1,971	1.44	4,077	3,261	1.25	5,619	4,473	1.26	8,526	6,582	1.30	6,966	5,925	1.18
Total Pacific	432	318	1.36	936	765	1.22	1,122	915	1.22	1,995	1,635	1.22	1,896	1,713	1.11
Total Asian							588	366	1.60	1,500	1,110	1.35	1,827	1,581	1.15
nMPA†	48,228	50,400	0.96	57,666	59,466	0.97	69,600	70,230	0.99	83,808	78,636	1.07	72,372	70,464	1.03
Missing & nMPA	49,222	50,406	0.98	58,332	59,484	0.98	69,873	71,406	0.98	84,736	87,141	0.97	73,272	74,646	0.98
Missing Ethnicity							273	1,176	0.23	918	8,502	0.11	900	4,179	0.22

* Each cohort includes all weighted HPL cancer registrations in this period compared back to their linked census record. † nMPA is a residual category of people who do not report affiliation with Māori and/or Pacific and /or Asian ethnic groups ^ Ratios are the figures that need to be multiplied to NZCR counts to give a 'correct' estimate of the 'gold standard' census Māori count figures. Formula = number of people with specific ethnic group on census/number of people with specific ethnic group on Cancer Register Note: numbers of cancers in this table were random rounded in accordance with Statistics New Zealand policy. Results from cells with very small numbers have been suppressed. The 1986 Total Asian ratio was 360/114 or 3.14. It was not placed in the table as the Cancer Registration Form send by hospitals to the Cancer register did not contain an Asian ethnic group option over most of this time period.

Table 26 Misclassification ratios for all cancers by total ethnicity, age group and cohort 1981-2004

	1981-86			1986-91			1991-96			1996-01			2001-04		
Age Group	Census Ethn	First Cancer Eth	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio
Both Sexes All															
Mäori															
All Ages	2,829	1,971	1.44	4,077	3,261	1.25	5,619	4,473	1.26	8,526	6,582	1.30	6,966	5,925	1.18
0- 4 yrs	42	24	1.64	42	30	1.38	42	33	1.24	66	57	1.20	42	36	1.19
5- 9 yrs	30	24	1.24	27	21	1.37	33	30	1.13	57	39	1.57	21	21	1.10
10-14 yrs	36	24	1.57	66	42	1.49	63	45	1.41	75	48	1.54	45	33	1.35
15-19 yrs	90	66	1.38	126	99	1.31	252	177	1.42	324	222	1.45	258	219	1.19
20-24 yrs	231	150	1.53	258	210	1.23	366	282	1.29	633	456	1.39	441	339	1.30
25-29 yrs	231	141	1.62	357	279	1.28	486	387	1.25	639	537	1.19	528	441	1.20
30-34 yrs	216	132	1.65	312	258	1.21	462	348	1.32	651	507	1.28	432	372	1.16
35-39 yrs	186	147	1.27	327	249	1.32	384	309	1.24	660	507	1.30	480	414	1.15
40-44 yrs	207	159	1.31	327	264	1.23	447	354	1.26	552	438	1.26	489	414	1.18
45-49 yrs	285	219	1.29	375	327	1.15	477	372	1.28	729	540	1.35	531	447	1.19
50-54 yrs	282	195	1.45	396	351	1.13	528	426	1.24	729	579	1.26	627	522	1.20
55-59 yrs	264	204	1.29	420	312	1.36	579	513	1.13	894	720	1.24	693	618	1.12
60-64 yrs	270	195	1.37	369	300	1.23	507	405	1.25	828	660	1.26	813	708	1.15
65-69 yrs	201	135	1.48	294	264	1.11	438	360	1.21	744	591	1.26	684	567	1.21
70-74 yrs	129	90	1.41	216	162	1.33	282	216	1.31	501	357	1.41	480	420	1.15
75-79 yrs	66	33	2.09	108	66	1.63	177	129	1.35	249	183	1.35	246	222	1.10
>=80 years	63	24	2.48	60	39	1.56	111	87	1.24	198	141	1.42	153	129	1.19
Total Pacific People															
All Ages	432	318	1.36	936	765	1.22	1,122	915	1.22	1,995	1,635	1.22	1,896	1,713	1.11

	1981-86			1986-91			1991-96			1996-01			2001-04		
Age Group	Census Ethn	First Cancer Eth	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio
0- 4 yrs	.	.	.	.	.	.	24	18	1.33	36	24	1.52	.	.	.
5- 9 yrs	.	.	.	21	15	1.57	.	.	.	.	.	.	21	21	1.05
10-14 yrs	.	.	.	33	24	1.32	24	18	1.53	30	18	1.61	.	.	.
15-19 yrs	.	.	.	42	21	2.05	57	33	1.65	54	36	1.54	69	48	1.46
20-24 yrs	24	9	2.88	48	15	2.94	96	60	1.58	99	57	1.78	66	42	1.59
25-29 yrs	42	27	1.52	69	57	1.21	87	63	1.44	138	96	1.45	105	75	1.38
30-34 yrs	45	39	1.15	81	75	1.07	87	75	1.17	126	99	1.25	123	111	1.12
35-39 yrs	42	30	1.35	93	69	1.35	81	78	1.01	147	117	1.27	117	102	1.15
40-44 yrs	36	27	1.41	96	75	1.27	72	57	1.22	147	129	1.15	144	132	1.11
45-49 yrs	39	36	1.14	75	66	1.12	78	66	1.18	156	135	1.14	162	141	1.16
50-54 yrs	45	36	1.35	75	75	1.00	105	102	1.03	186	174	1.08	153	150	1.02
55-59 yrs	27	24	1.17	87	84	1.05	108	84	1.28	234	207	1.13	189	177	1.07
60-64 yrs	36	27	1.33	78	75	1.07	96	87	1.08	177	150	1.18	207	198	1.06
65-69 yrs	30	24	1.25	51	48	1.15	93	84	1.13	189	162	1.17	198	198	1.00
70-74 yrs	.	.	.	42	33	1.21	63	48	1.33	132	108	1.20	150	144	1.03
75-79 yrs	.	.	.	24	21	1.19	24	21	1.20	84	69	1.20	84	84	1.01
>=80 years	.	.	.	.	.	.	.	.	.	54	42	1.32	66	66	0.98
Total Asian															
All Ages	.	.	.	360	114	3.14	588	366	1.60	1,500	1,110	1.35	1,827	1,581	1.15
15-19 yrs	.	.	.	.	.	.	24	12	1.83	33	15	2.13	33	21	1.55
20-24 yrs	.	.	.	.	.	.	27	12	2.42	69	24	2.65	75	45	1.61
25-29 yrs	.	.	.	27	9	3.38	45	21	2.24	99	60	1.62	105	81	1.27
30-34 yrs	.	.	.	.	.	.	54	33	1.66	108	78	1.39	123	105	1.16
35-39 yrs	.	.	.	27	9	2.70	63	27	2.37	150	96	1.56	171	144	1.17

	1981-86			1986-91			1991-96			1996-01			2001-04		
Age Group	Census Ethn	First Cancer Eth	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio
40-44 yrs	.	.	.	.	.	.	57	33	1.68	174	132	1.32	174	156	1.10
45-49 yrs	.	.	.	.	.	.	57	33	1.84	195	135	1.44	246	201	1.22
50-54 yrs	.	.	.	45	18	2.59	51	36	1.38	129	93	1.36	207	189	1.10
55-59 yrs	.	.	.	30	15	2.23	51	39	1.33	141	108	1.30	159	144	1.11
60-64 yrs	.	.	.	30	15	2.29	36	24	1.38	93	81	1.12	153	144	1.06
65-69 yrs	.	.	.	.	.	.	33	21	1.35	102	96	1.07	129	114	1.12
70-74 yrs	.	.	.	21	9	2.63	30	24	1.12	78	57	1.41	99	93	1.08
75-79 yrs	.	.	.	.	.	.	27	21	1.17	45	42	1.05	66	63	1.10
>=80 years	.	.	.	.	.	.	.	.	.	63	54	1.15	48	45	1.04
nonMPA (European/Other)															
All Ages	48,228	50,400	0.96	57,666	59,466	0.97	69,600	70,230	0.99	83,808	78,636	1.07	72,372	70,464	1.03
0- 4 yrs	159	177	0.90	165	174	0.95	159	174	0.92	171	183	0.93	117	129	0.92
5- 9 yrs	99	117	0.87	120	132	0.92	108	111	0.97	102	108	0.95	87	87	0.98
10-14 yrs	207	222	0.92	210	240	0.88	219	237	0.91	228	234	0.97	150	159	0.94
15-19 yrs	444	480	0.92	720	780	0.92	888	987	0.90	1,023	1,056	0.97	834	867	0.96
20-24 yrs	912	1,008	0.90	1,407	1,497	0.94	1,566	1,692	0.93	2,166	2,307	0.94	1,656	1,740	0.95
25-29 yrs	1,449	1,575	0.92	2,070	2,190	0.95	1,818	1,953	0.93	2,631	2,634	1.00	2,133	2,196	0.97
30-34 yrs	1,707	1,827	0.93	2,070	2,166	0.96	2,103	2,223	0.95	2,601	2,574	1.01	2,265	2,283	0.99
35-39 yrs	1,701	1,785	0.95	2,445	2,571	0.95	2,325	2,403	0.97	2,904	2,808	1.03	2,394	2,370	1.01
40-44 yrs	1,911	2,007	0.95	2,454	2,553	0.96	3,048	3,111	0.98	3,474	3,186	1.09	2,979	2,904	1.03
45-49 yrs	2,595	2,697	0.96	3,072	3,159	0.97	3,711	3,741	0.99	5,124	4,545	1.13	4,143	3,885	1.07
50-54 yrs	3,669	3,813	0.96	3,579	3,669	0.97	4,416	4,401	1.00	5,883	5,202	1.13	5,649	5,292	1.07
55-59 yrs	5,394	5,541	0.97	5,592	5,775	0.97	5,961	5,919	1.01	7,512	6,762	1.11	6,693	6,249	1.07
60-64 yrs	5,961	6,168	0.97	7,263	7,413	0.98	8,334	8,313	1.00	8,745	7,902	1.11	7,830	7,422	1.06

	1981-86			1986-91			1991-96			1996-01			2001-04		
Age Group	Census Ethn	First Cancer Eth	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio
65-69 yrs	6,807	7,053	0.97	7,422	7,563	0.98	10,227	10,176	1.00	11,583	10,698	1.08	8,739	8,430	1.04
70-74 yrs	6,543	6,822	0.96	7,689	7,875	0.98	9,147	9,159	1.00	11,655	10,974	1.06	9,420	9,216	1.02
75-79 yrs	4,491	4,713	0.95	5,976	6,141	0.97	7,737	7,755	1.00	8,580	8,238	1.04	8,352	8,241	1.01
>=80 years	4,173	4,395	0.95	5,415	5,565	0.97	7,827	7,878	0.99	9,423	9,231	1.02	8,928	8,991	0.99
nonMPA&Miss															
All Ages	49,266	50,406	0.98	58,332	59,484	0.98	69,873	71,406	0.98	84,726	87,141	0.97	73,272	74,646	0.98
0- 4 yrs	162	180	0.91	168	174	0.97	159	174	0.91	174	186	0.94	117	126	0.92
5- 9 yrs	99	117	0.87	123	132	0.94	108	114	0.96	102	117	0.90	90	90	0.98
10-14 yrs	207	225	0.92	213	240	0.89	219	243	0.91	228	261	0.88	150	162	0.91
15-19 yrs	453	480	0.94	732	783	0.93	891	993	0.9	1,026	1,146	0.90	843	906	0.93
20-24 yrs	921	1,011	0.91	1,416	1,497	0.95	1,575	1,704	0.92	2,202	2,442	0.90	1,671	1,809	0.92
25-29 yrs	1,467	1,575	0.93	2,088	2,187	0.95	1,824	1,971	0.92	2,646	2,796	0.95	2,157	2,289	0.94
30-34 yrs	1,716	1,824	0.94	2,082	2,166	0.96	2,109	2,253	0.94	2,625	2,802	0.94	2,277	2,355	0.97
35-39 yrs	1,722	1,788	0.96	2,454	2,571	0.95	2,334	2,439	0.96	2,928	3,129	0.94	2,412	2,511	0.96
40-44 yrs	1,926	2,007	0.96	2,463	2,559	0.96	3,054	3,174	0.96	3,510	3,672	0.96	3,000	3,096	0.97
45-49 yrs	2,610	2,694	0.97	3,084	3,159	0.98	3,717	3,852	0.96	5,163	5,406	0.96	4,182	4,320	0.97
50-54 yrs	3,696	3,816	0.97	3,600	3,672	0.98	4,428	4,545	0.97	5,943	6,126	0.97	5,715	5,841	0.98
55-59 yrs	5,460	5,544	0.99	5,646	5,775	0.98	5,979	6,081	0.98	7,566	7,770	0.97	6,756	6,852	0.99
60-64 yrs	6,072	6,168	0.98	7,329	7,416	0.99	8,361	8,478	0.99	8,817	9,006	0.98	7,902	8,016	0.99
65-69 yrs	6,975	7,053	0.99	7,515	7,563	0.99	10,260	10,359	0.99	11,706	11,871	0.99	8,829	8,958	0.99
70-74 yrs	6,768	6,822	0.99	7,806	7,878	0.99	9,192	9,273	0.99	11,784	11,961	0.99	9,552	9,621	0.99
75-79 yrs	4,668	4,713	0.99	6,087	6,144	0.99	7,788	7,839	0.99	8,712	8,784	0.99	8,490	8,517	1.00
>=80 years	4,347	4,395	0.99	5,526	5,568	0.99	7,884	7,917	1.00	9,591	9,663	0.99	9,135	9,159	1.00
Missing Ethnicity															

	1981-86			1986-91			1991-96			1996-01			2001-04		
Age Group	Census Ethn	First Cancer Eth	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio	Census Ethn	First Cancer Ethn	Census / 1st Cancer Ratio
All Ages	.	.	.	.	.	.	273	1,176	0.23	918	8,502	0.11	900	4,179	0.22
15-19 yrs	.	.	.	.	.	.	.	.	.	.	.	.	9	36	0.24
20-24 yrs	.	.	.	.	.	.	.	.	.	36	135	0.26	12	66	0.19
25-29 yrs	.	.	.	.	.	.	.	.	.	18	159	0.11	27	96	0.27
30-34 yrs	.	.	.	.	.	.	.	.	.	27	228	0.12	12	72	0.18
35-39 yrs	.	.	.	.	.	.	.	.	.	24	324	0.07	18	141	0.13
40-44 yrs	.	.	.	.	.	.	.	.	.	36	483	0.07	21	192	0.11
45-49 yrs	.	.	.	.	.	.	9	114	0.06	39	864	0.04	42	438	0.09
50-54 yrs	.	.	.	.	.	.	12	144	0.08	57	921	0.06	66	555	0.12
55-59 yrs	.	.	.	.	.	.	18	159	0.10	57	1,008	0.05	60	606	0.10
60-64 yrs	.	.	.	.	.	.	21	165	0.14	69	1,107	0.06	69	597	0.12
65-69 yrs	.	.	.	.	.	.	33	180	0.19	123	1,173	0.10	90	525	0.17
70-74 yrs	.	.	.	.	.	.	42	114	0.38	129	987	0.13	129	405	0.32
75-79 yrs	.	.	.	.	.	.	51	84	0.58	135	546	0.24	141	276	0.51
>=80 years	.	.	.	.	.	.	57	39	1.43	165	435	0.38	207	171	1.21

Table 27 Misclassification ratios for all cancers by total ethnicity, by sex and cohort 1981-2004

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Males															
All															
Mäori	1,086	771	1.41	1,335	1,050	1.27	1,959	1,563	1.25	2,910	2,175	1.34	2,532	2,136	1.19
Total Pacific	162	129	1.25	309	264	1.17	438	354	1.23	765	648	1.18	702	642	1.09
Total Asian	.	.	.	168	69	2.49	240	171	1.41	516	393	1.32	573	519	1.11
nonMPA	23,163	23,994	0.97	26,223	26,889	0.98	33,414	33,420	1.00	40,599	37,185	1.09	34,521	32,970	1.05
nonMPA&Miss	23,574	23,994	0.98	26,487	26,892	0.98	33,525	34,062	0.98	40,989	41,871	0.98	34,941	35,421	0.99
Missing Ethnicity							111	639	0.18	390	4,683	0.08	420	2,451	0.17

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Females															
All															
Mäori	1,746	1,203	1.45	2,742	2,211	1.24	3,660	2,907	1.26	5,616	4,407	1.27	4,434	3,789	1.17
Total Pacific	270	189	1.44	624	504	1.25	684	561	1.22	1,236	987	1.25	1,197	1,071	1.12
Total Asian	.	.	.	189	45	4.20	348	198	1.77	984	717	1.37	1,254	1,062	1.18
nonMPA	25,065	26,406	0.95	31,443	32,580	0.97	36,189	36,807	0.98	43,209	41,451	1.04	37,851	37,497	1.01
nonMPA&Miss	25,695	26,412	0.97	31,848	32,589	0.98	36,345	37,341	0.97	43,740	45,270	0.97	38,331	39,225	0.98
Missing Ethnicity	.	.	.	.	.	.	159	534	0.30	531	3,819	0.14	480	1,728	0.28

Table 28 Misclassification ratios for all cancers by total ethnicity, NZDep and cohort 1981-2004

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Both Sexes															
Missing Dep															
nonMPA	18	24	0.78	15	21	0.74	30	33	0.97	42	42	0.98	30	33	0.91
nonMPA&Miss	18	24	0.78	.	.	.	30	30	0.97	42	45	0.91	30	33	0.91
NZDep Decile 1															
Mäori	72	36	1.87	66	42	1.68	87	57	1.57	195	108	1.81	150	111	1.36
Total Pacific	.	.	.	.	.	.	.	.	.	45	24	1.69	33	42	0.85

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Total Asian	.	.	.	33	15	2.69	54	36	1.42	153	135	1.13	201	174	1.16
nonMPA	4,056	4,182	0.97	4,743	4,830	0.98	6,267	6,165	1.02	8,601	7,572	1.14	8,040	7,500	1.07
nonMPA&Miss	4,140	4,182	0.99	4,788	4,830	0.99	6,285	6,339	0.99	8,661	8,778	0.99	8,100	8,160	0.99
Missing Ethn	.	.	.	.	.	.	18	174	0.10	57	1,203	0.05	63	660	0.10
NZDep Decile 2															
Mäori	93	30	3.10	102	69	1.46	141	72	1.99	312	189	1.65	231	159	1.45
Total Pacific	.	.	.	33	24	1.46	24	12	1.85	57	42	1.41	54	42	1.26
Total Asian	.	.	.	30	12	2.50	66	36	1.78	165	141	1.16	189	174	1.08
nonMPA	4,134	4,290	0.96	4,869	4,971	0.98	6,312	6,294	1.00	8,493	7,614	1.12	7,992	7,530	1.06
nonMPA&Miss	4,212	4,290	0.98	4,914	4,974	0.99	6,336	6,444	0.98	8,571	8,724	0.98	8,055	8,148	0.99
Missing Ethn	.	.	.	.	.	.	24	153	0.16	78	1,116	0.07	63	618	0.10
NZDep Decile 3															
Mäori	99	69	1.41	135	96	1.45	183	120	1.52	363	204	1.78	303	216	1.41
Total Pacific	27	24	1.22	33	24	1.36	51	42	1.27	96	84	1.14	72	63	1.13
Total Asian	.	.	.	36	12	3.00	57	42	1.36	153	111	1.38	195	162	1.18
nonMPA	4,575	4,707	0.97	5,469	5,580	0.98	6,927	6,915	1.00	9,033	8,307	1.09	8,247	7,860	1.05
nonMPA&Miss	4,653	4,710	0.99	5,517	5,583	0.99	6,957	7,041	0.99	9,126	9,324	0.98	8,325	8,451	0.98
Missing Ethn	.	.	.	.	.	.	30	126	0.23	93	1,020	0.09	75	591	0.13
NZDep Decile 4															
Mäori	129	75	1.72	201	141	1.44	234	168	1.39	432	282	1.53	345	255	1.34
Total Pacific	21	15	1.69	69	51	1.31	45	42	1.07	93	75	1.23	87	72	1.21
Total Asian	.	.	.	.	.	.	48	36	1.33	141	114	1.27	177	156	1.15
nonMPA	4,623	4,794	0.96	5,724	5,865	0.98	7,074	7,062	1.00	8,919	8,217	1.09	7,833	7,563	1.04
nonMPA&Miss	4,710	4,797	0.98	5,781	5,865	0.99	7,104	7,185	0.99	9,000	9,174	0.98	7,917	8,037	0.99

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Missing Ethn	.	.	.	.	.	.	30	123	0.24	81	960	0.09	84	471	0.18
NZDep Decile 5															
Mäori	162	102	1.59	231	171	1.35	273	225	1.22	537	369	1.45	450	360	1.25
Total Pacific	30	18	1.71	54	51	1.08	66	51	1.33	114	84	1.31	111	99	1.11
Total Asian	.	.	.	30	9	3.33	60	33	1.74	138	99	1.41	198	156	1.26
nonMPA	5,244	5,436	0.96	6,333	6,492	0.98	7,800	7,806	1.00	9,531	8,874	1.07	8,109	7,857	1.03
nonMPA&Miss	5,349	5,439	0.98	6,414	6,495	0.99	7,836	7,923	0.99	9,615	9,834	0.98	8,196	8,331	0.98
Missing Ethn	.	.	.	.	.	.	33	120	0.29	84	957	0.09	84	477	0.18
NZDep Decile 6															
Mäori	240	159	1.53	315	222	1.42	408	285	1.44	627	438	1.43	576	450	1.27
Total Pacific	36	24	1.48	96	66	1.45	72	60	1.16	141	117	1.19	102	78	1.32
Total Asian	.	.	.	42	15	2.87	72	42	1.69	162	111	1.44	189	171	1.10
nonMPA	5,304	5,514	0.96	6,456	6,678	0.97	7,908	7,980	0.99	9,288	8,805	1.05	7,851	7,704	1.02
nonMPA&Miss	5,403	5,517	0.98	6,534	6,678	0.98	7,938	8,103	0.98	9,399	9,627	0.98	7,959	8,115	0.98
Missing Ethn	.	.	.	.	.	.	33	120	0.26	108	819	0.13	108	408	0.26
NZDep Decile 7															
Mäori	294	204	1.45	411	324	1.26	588	438	1.35	879	669	1.31	732	582	1.26
Total Pacific	36	27	1.42	84	75	1.11	96	81	1.21	147	126	1.18	159	138	1.15
Total Asian	.	.	.	.	.	.	57	24	2.15	159	108	1.49	180	156	1.15
nonMPA	5,571	5,832	0.96	6,681	6,882	0.97	7,956	8,073	0.99	9,111	8,700	1.05	7,515	7,494	1.00
nonMPA&Miss	5,700	5,832	0.98	6,777	6,885	0.98	7,983	8,172	0.98	9,222	9,474	0.97	7,623	7,809	0.98
Missing Ethn	.	.	.	.	.	.	24	99	0.25	108	774	0.14	108	318	0.34
NZDep Decile 8															
Mäori	390	276	1.42	537	426	1.27	789	630	1.25	1,080	828	1.30	933	804	1.16

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Total Pacific	48	36	1.26	93	75	1.24	126	93	1.35	213	162	1.32	258	240	1.07
Total Asian	.	.	.	57	18	3.17	66	45	1.52	165	114	1.43	195	168	1.16
nonMPA	5,493	5,766	0.95	6,723	6,960	0.97	7,758	7,890	0.98	8,586	8,361	1.03	6,978	6,966	1.00
nonMPA&Miss	5,619	5,766	0.97	6,795	6,957	0.98	7,788	7,998	0.97	8,679	8,991	0.97	7,095	7,257	0.98
Missing Ethn	.	.	.	.	.	.	27	108	0.27	93	630	0.15	117	291	0.40
NZDep Decile 9															
Mäori	528	384	1.37	810	675	1.20	1,140	921	1.24	1,599	1,311	1.22	1,287	1,137	1.13
Total Pacific	81	69	1.17	165	138	1.21	213	174	1.24	417	333	1.25	372	342	1.08
Total Asian	.	.	.	36	12	3.08	60	36	1.61	171	114	1.48	174	150	1.15
nonMPA	5,124	5,448	0.94	6,060	6,312	0.96	6,681	6,894	0.97	7,326	7,221	1.01	5,919	5,964	0.99
nonMPA&Miss	5,277	5,448	0.97	6,141	6,315	0.97	6,708	6,981	0.96	7,446	7,809	0.95	6,015	6,192	0.97
Missing Ethn	.	.	.	.	.	.	27	87	0.31	120	588	0.21	96	228	0.41
NZDep Decile 10															
Mäori	810	630	1.28	1,257	1,092	1.15	1,761	1,551	1.14	2,487	2,163	1.15	1,941	1,836	1.06
Total Pacific	123	90	1.37	294	246	1.20	408	351	1.17	681	588	1.16	645	594	1.09
Total Asian	.	.	.	42	15	3.31	51	33	1.61	96	63	1.52	135	117	1.17
nonMPA	4,083	4,401	0.93	4,587	4,881	0.94	4,890	5,121	0.95	4,875	4,923	0.99	3,861	3,987	0.97
nonMPA&Miss	4,185	4,401	0.95	4,656	4,878	0.95	4,911	5,181	0.95	4,971	5,358	0.93	3,963	4,110	0.83
Missing Ethn	.	.	.	.	.	.	21	60	0.38	96	435	0.22	102	123	0.83

Table 29 Misclassification ratios for all cancers by total ethnicity, Regional Health Authority (RHA) and cohort 1981-2004

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Both Sexes															
Northern															
Mäori	1,059	699	1.51	1,350	1,083	1.24	1,791	1,455	1.23	2,805	2,196	1.28	2,211	1,890	1.17
Total Pacific	312	240	1.31	618	528	1.17	831	690	1.20	1,374	1,164	1.18	1,305	1,206	1.09
Total Asian	.	.	.	162	66	2.44	321	219	1.46	924	699	1.32	1,194	1,044	1.15
nonMPA	14,697	15,513	0.95	16,551	17,133	0.97	20,601	20,730	0.99	26,088	23,895	1.09	22,527	21,759	1.04
nonMPA&Miss	15,012	15,513	0.97	16,740	17,139	0.98	20,688	21,240	0.97	26,397	27,300	0.97	22,884	23,403	0.98
Missing Ethn	.	.	.	.	.	.	87	507	0.17	309	3,405	0.09	357	1,644	0.22
Midland															
Mäori	870	726	1.20	1,392	1,209	1.15	2,043	1,779	1.15	3,009	2,550	1.18	2,376	2,166	1.10
Total Pacific	27	12	2.00	75	42	1.93	84	57	1.48	165	78	2.10	162	123	1.31
Total Asian	.	.	.	42	12	3.33	54	30	1.96	102	66	1.56	108	84	1.27
nonMPA	7,998	8,289	0.96	10,416	10,755	0.97	13,287	13,500	0.98	15,504	14,736	1.05	14,361	13,932	1.03
nonMPA&Miss	8,115	8,292	0.98	10,527	10,758	0.98	13,347	13,653	0.98	15,687	16,200	0.97	14,541	14,799	0.98
Missing Ethn	.	.	.	.	.	.	60	153	0.39	180	1,464	0.12	180	867	0.21
Central															
Mäori	666	453	1.47	981	780	1.26	1,281	945	1.35	1,893	1,368	1.38	1,686	1,458	1.16
Total Pacific	75	54	1.43	180	156	1.15	159	135	1.18	336	294	1.14	324	303	1.06
Total Asian	.	.	.	120	27	4.44	138	75	1.86	303	231	1.31	318	282	1.13
nonMPA	12,504	13,092	0.96	14,949	15,423	0.97	17,379	17,553	0.99	20,802	19,557	1.06	17,919	17,487	1.02
nonMPA&Miss	12,804	13,092	0.98	15,123	15,432	0.98	17,445	17,853	0.98	21,021	21,606	0.97	18,084	18,357	0.99
Missing Ethn	.	.	.	.	.	.	66	300	0.22	219	2,046	0.11	165	870	0.19
Southern															

Total Ethnicity	1981-86			1986-91			1991-96			1996-01			2001-04		
	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Mäori	237	93	2.52	357	192	1.87	507	288	1.75	825	465	1.77	690	414	1.67
Total Pacific	.	.	.	60	42	1.42	45	33	1.42	120	96	1.26	102	81	1.24
Total Asian	.	.	.	36	9	4.63	75	48	1.63	171	111	1.54	204	171	1.20
nonMPA	13,029	13,509	0.96	15,747	16,152	0.97	18,339	18,447	0.99	21,411	20,451	1.05	17,565	17,286	1.02
nonMPA&Miss	13,338	13,509	0.99	15,945	16,155	0.99	18,396	18,657	0.99	21,624	22,038	0.98	17,763	18,084	0.98
Missing Ethn	.	.	.	.	.	.	60	210	0.29	213	1,587	0.13	201	792	0.25

Table 30 Misclassification ratios for all cancers for Māori and non-Māori by DHB and cohort 1981-2004

Total Ethnicity	1981-86			1986-91			1991-96			1996-01			2001-04		
	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Both Sexes															
Northland															
Mäori	240	159	1.50	306	288	1.07	540	483	1.11	837	768	1.10	657	621	1.05
Non-Mäori	1,284	1,380	0.93	1,698	1,749	0.97	2,316	2,364	0.98	3,069	2,967	1.04	2,946	2,964	0.99
Waitemata															
Mäori	216	129	1.68	249	165	1.49	360	258	1.40	561	330	1.71	420	321	1.31
Non-Mäori	4,782	4,971	0.96	5,748	5,868	0.98	7,428	7,371	1.01	10,263	9,057	1.13	9,060	8,508	1.07
Auckland															
Mäori	303	192	1.58	348	279	1.25	333	264	1.27	537	396	1.35	423	330	1.29
Non-Mäori	6,012	6,240	0.96	6,045	6,183	0.98	6,960	6,861	1.01	8,445	7,797	1.08	7,050	6,795	1.04

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Counties Manukau															
Mäori	300	216	1.38	444	351	1.26	561	450	1.24	864	705	1.23	714	618	1.16
Non-Mäori	3,006	3,168	0.95	3,795	3,927	0.97	5,022	5,043	1.00	6,456	5,925	1.09	5,886	5,709	1.03
Waikato															
Mäori	309	243	1.26	465	369	1.27	690	579	1.19	972	798	1.22	759	678	1.12
Non-Mäori	3,417	3,534	0.97	4,551	4,698	0.97	5,733	5,805	0.99	6,279	5,877	1.07	5,754	5,652	1.02
Lakes															
Mäori	201	162	1.23	270	240	1.12	429	378	1.14	546	495	1.10	486	465	1.04
Non-Mäori	819	870	0.94	1,029	1,071	0.96	1,368	1,410	0.97	1,659	1,599	1.04	1,593	1,545	1.03
Bay of Plenty															
Mäori	174	150	1.15	339	321	1.06	501	426	1.17	789	663	1.19	636	573	1.11
Non-Mäori	1,629	1,677	0.97	2,127	2,166	0.98	3,243	3,276	0.99	4,032	3,759	1.07	4,059	3,813	1.06
Tairawhiti															
Mäori	123	105	1.17	189	171	1.10	249	243	1.02	447	411	1.08	309	291	1.05
Non-Mäori	483	504	0.96	672	693	0.97	699	696	1.00	1,023	996	1.03	690	690	1.00
Taranaki															
Mäori	69	66	1.05	129	108	1.18	174	153	1.15	255	180	1.42	192	156	1.23
Non-Mäori	1,695	1,716	0.99	2,136	2,178	0.98	2,373	2,394	0.99	2,709	2,649	1.02	2,502	2,427	1.03
Hawke's Bay															
Mäori	171	123	1.39	273	246	1.11	381	309	1.24	513	399	1.28	456	435	1.06
Non-Mäori	1,863	1,950	0.96	2,541	2,583	0.98	2,874	2,856	1.01	3,771	3,555	1.06	2,997	2,871	1.04
Whanganui															
Mäori	84	63	1.33	105	96	1.09	141	111	1.30	252	210	1.19	222	204	1.10
Non-Mäori	918	972	0.95	1,242	1,266	0.98	1,497	1,524	0.98	1,734	1,701	1.02	1,359	1,350	1.01
MidCentral															

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Mäori	126	63	1.97	186	156	1.19	273	201	1.37	369	252	1.46	306	228	1.34
Non-Mäori	2,442	2,547	0.96	2,937	3,006	0.98	3,543	3,576	0.99	3,819	3,642	1.05	3,303	3,165	1.04
Hutt Valley															
Mäori	108	72	1.50	114	96	1.18	126	93	1.33	219	189	1.17	198	180	1.09
Non-Mäori	2,154	2,238	0.96	2,577	2,631	0.98	2,376	2,376	1.00	2,910	2,733	1.07	2,565	2,505	1.02
Capital & Coast															
Mäori	114	90	1.28	198	123	1.64	216	147	1.48	291	177	1.64	270	222	1.21
Non-Mäori	3,327	3,447	0.97	3,480	3,588	0.97	4,104	4,122	1.00	4,947	4,353	1.14	4,368	4,230	1.03
Wairarapa															
Mäori	30	18	1.76	51	42	1.25	84	66	1.28	111	84	1.30	72	81	0.89
Non-Mäori	549	573	0.96	753	774	0.98	957	969	0.99	999	957	1.05	930	900	1.04
Nelson Marlborough															
Mäori	33	24	1.32	54	21	2.45	60	24	2.27	132	57	2.44	159	111	1.44
Non-Mäori	1,377	1,419	0.97	1,707	1,755	0.97	2,313	2,337	0.99	3,195	3,132	1.02	3,003	3,033	0.99
West Coast															
Mäori	.	.	.	.	.	.	30	15	1.87	57	39	1.55	36	21	1.73
Non-Mäori	516	546	0.95	783	795	0.98	774	783	0.99	909	909	1.00	660	675	0.98
Canterbury															
Mäori	99	39	2.53	180	99	1.83	234	111	2.05	375	216	1.73	375	219	1.72
Non-Mäori	6,687	6,924	0.97	8,202	8,379	0.98	9,210	9,201	1.00	11,598	10,755	1.08	9,594	9,273	1.03
South Canterbury															
Mäori	27	9	2.55	.	.	.	36	24	1.58	57	33	1.84	33	18	1.84
Non-Mäori	1,119	1,143	0.98	1,308	1,332	0.98	1,467	1,482	0.99	1,725	1,659	1.04	1,425	1,413	1.01
Otago															
Mäori	42	15	3.23	69	33	2.00	102	63	1.62	177	87	2.01	123	78	1.54

Total Ethnicity	1981-86			1986-91			1991-96			1996-01			2001-04		
	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Non-Mäori	3,417	3,519	0.97	3,921	3,996	0.98	4,869	4,890	1.00	5,043	4,956	1.02	4,263	4,251	1.00
Southland															
Mäori	57	27	2.11	78	42	1.86	105	72	1.46	153	93	1.68	117	75	1.59
Non-Mäori	1,332	1,389	0.96	1,632	1,695	0.96	2,136	2,169	0.98	2,403	2,379	1.01	1,914	1,923	0.99

Table 31 Misclassification ratios for all cancers by total ethnicity, rurality and cohort 1981-2004

Total Ethnicity	1981-86			1986-91			1991-96			1996-01			2001-04		
	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
Both Sexes															
Main Urban															
Mäori	1,848	1,260	1.47	2,463	1,911	1.29	3,183	2,523	1.26	4,974	3,690	1.35	4,188	3,534	1.19
Total Pacific	396	294	1.33	873	738	1.18	1,050	858	1.23	1,821	1,542	1.18	1,746	1,623	1.08
Total Asian	.	.	.	312	93	3.39	513	327	1.56	1,359	1,032	1.32	1,713	1,497	1.15
nonMPA	35,919	37,563	0.96	41,328	42,606	0.97	48,750	49,035	0.99	59,151	54,813	1.08	50,385	48,762	1.03
nonMPA&Miss	36,726	37,572	0.98	41,778	42,621	0.98	48,954	49,947	0.98	59,778	61,455	0.97	51,012	51,939	0.98
Missing Ethn	.	.	.	.	.	.	201	915	0.22	627	6,642	0.09	624	3,174	0.20
Secondary Urban Area															
Mäori	183	123	1.51	324	261	1.25	417	294	1.41	633	480	1.32	501	435	1.15
Total Pacific	.	.	.	33	15	2.06	33	33	1.09	48	33	1.48	45	30	1.45
Total Asian	.	.	.	.	.	.	33	21	1.65	45	27	1.76	39	30	1.34

	1981-86			1986-91			1991-96			1996-01			2001-04		
Total Ethnicity	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio	Census Eth	First Cancer Eth	Census / 1st Cancer Ratio
nonMPA	3,930	4,095	0.96	4,965	5,115	0.97	6,009	6,102	0.98	6,648	6,468	1.03	5,778	5,694	1.01
nonMPA&Miss	4,023	4,095	0.98	5,034	5,115	0.98	6,021	6,153	0.98	6,723	6,885	0.98	5,835	5,919	0.99
Missing Ethn	.	.	.	.	.	.	12	54	0.25	72	417	0.17	57	225	0.25
Minor Urban Area															
Mäori	453	321	1.41	714	600	1.19	1,095	879	1.25	1,461	1,179	1.24	1,137	972	1.17
Total Pacific	.	.	.	.	.	.	21	18	1.18	63	27	2.25	66	36	1.76
Total Asian	.	.	.	27	9	2.55	21	15	1.53	54	27	1.86	45	36	1.28
nonMPA	5,205	5,436	0.96	6,627	6,858	0.97	8,214	8,373	0.98	9,255	8,997	1.03	7,761	7,740	1.00
nonMPA&Miss	5,292	5,436	0.97	6,723	6,858	0.98	8,241	8,469	0.97	9,390	9,705	0.97	7,881	8,073	0.98
Missing Ethn	.	.	.	.	.	.	24	93	0.28	135	708	0.19	117	333	0.35
Rural															
Mäori	345	267	1.29	573	495	1.16	930	774	1.20	1,458	1,230	1.19	1,137	987	1.16
Total Pacific	.	.	.	21	12	1.75	.	.	.	63	30	2.17	36	21	1.73
Total Asian	.	.	.	.	.	.	.	.	.	45	21	2.00	30	21	1.38
nonMPA	3,171	3,306	0.96	4,746	4,887	0.97	6,627	6,720	0.99	8,754	8,361	1.05	8,442	8,268	1.02
nonMPA&Miss	3,222	3306	0.97	4797	4887	0.98	6657	6834	0.97	8841	9096	0.97	8547	8715	0.98
Missing Ethn	.	.	.	.	.	.	33	11	0.27	87	735	0.12	102	447	0.23

Linkage bias dataset and calculation of linkage weights

This section considers the issue of linkage bias in more detail. Linkage bias may arise if the proportion of cancer records linked to the census varies by sociodemographic factors.

The process of anonymous and probabilistic record linkage is imperfect, with 26.8%, 22.9%, 20.8%, 20.3% and 18.3% of records being unable to be linked respectively in the 1981, 1986, 1991, 1996 and 2001 cohorts.

If the probability of linkage varies by factors of interest (e.g. age, ethnicity, socio-economic position) then future rate ratio (and definitely rate difference) estimates of association between these factors and cancer in cohort analyses will be biased. Incomplete linkage between census and cancer means that some members of the census cohort are misclassified as cancer free when in reality they have developed cancer.

In all NZCMS cohorts, when the mortality and census records were stratified by demographic characteristics (age, sex and ethnicity), geographical distribution (rural/urban and Regional Health Authority), socioeconomic measures (NZ Deprivation Index), time following census and the level of mobility in the area unit, the proportion of mortality records linked varied by strata (i.e. linkage bias was present). In order to compensate for linkage bias, records in the NZCMS were weighted. In epidemiological terms, the weighting adjusts for misclassification of the mortality outcome in cohort analyses (Fawcett, Blakely et al. 2002; Fawcett, Atkinson et al. 2008).

This section documents the investigations into linkage bias in CancerTrends, the methods used to weight for the bias and the results of these weights.

1.8 Linkage bias dataset

A detailed description of the variables in the linkage bias dataset is in Appendix 6. In summary this dataset contains the demographic variables from the health datasets,

plus details of up to four different cancers the person may have had in this time period, plus a flag of whether the record was linked or not.

1.9 Linkage bias in CancerTrends

We followed our previous linkage bias exercises in the NZCMS, and the following guiding principles as below, in the calculation of linkage weights:

- We tried to have strata levels as detailed as possible but needed to combine some due to small numbers.
- Sex was always treated as separate strata, it was never combined.
- Aggregation was necessary to maintain numbers – every strata needed some linked records, it could not just have non-linked records.
- We used people with cancers for the strata and weights, not separate cancer records (people could have up to 4 cancers per cohort). They were linked independently of what cancers they had.
- Our principle was to try and have consistent strata and regimes across all 5 bias datasets with only minor differences between groupings.
- 0 – 14 year olds and 15 – 24 year olds were grouped into 0 – 24 year olds, because for all cohorts there were small numbers of cancers in these age groups. However, in retrospect there was good linkage in 0 – 14 year olds but relatively poor linkages in 15 – 24 year olds, meaning that it might have been better to keep these age groups separate. Therefore, there is a need to be cautious analysing and interpreting 0 - 24 year olds. If we conduct specific analyses for 15-24 year olds in the future, it may pay to make a new weight to use in analyses for this group.
- We used regression analyses separately for each of the 5 bias datasets to identify strata for collapsing – first including all missing data (as separate class or level) and then restricting to non-missing Territorial Authority, NZDep and ethnicity. Results of these regressions are in Table 116 and Table 117.
    - Regression was logistic using Proc Genmod in SAS.
    - We save the results of the Type I and Type III tests and made one dataset from all 10 regressions (5 years * 2).
    - Saved all parameter estimates into one dataset.
    - Looked at patterns over all analyses to determine which variables and levels were important or which ones could be combined.
    - Variables in regression were

- cenyear (always the same within each dataset), e.g. 1981, 1986, 1991, 1996, 2001)
- sex (male=1, female=2)
- Age (approximately ten year categories informat iageity. 0=' 0-14 yrs', 15='15-24 yrs', 25='25-29 yrs', 30='30-34 yrs', 35='35-39 yrs', 40='40-49 yrs', 50='50-59 yrs', 60='60-69 yrs', 70='70-79 yrs', 80='>=80 years')
- Maori on C1 (Maori ethnicity on first cancer diagnosis)
- Pacific on C1 (Pacific ethnicity on first cancer diagnosis)
- Asian on C1 (Asian ethnicity on first cancer diagnosis)
- nonMPA on C1 (non-Maori non-Pacific non-Asian ethnicity on first cancer diagnosis)
- Time since census (Number of months after census before first cancer (made negative so that the date closest to census is the reference group) -999='Missing', 0='Dates -ve, 0- 6 mths', -7=' 7-12 mths', -13='13-18 mths', -19='19-24 mths', -25='25-30 mths', -31='31-36 mths', -37='37-42 mths', -43='43-high mths')
- Territorial authority (Invercargill made reference group)
- NZ Deprivation index (3 groups Deciles 1–6 (ref), Deciles 7-8, Deciles 9-10). Had to use NZDep 2001 as the record linkage used the base 2001 versions of meshblock and area unit.
- Rurality (-1='Main Urban' (ref), -2='Secondary Urban Area', -3='Minor Urban Area', -7,-70='NonUrban or Missing') Also needed to use Rurality 2001 as this was the geocode base used for linkage.
- Dependent variable was linkage (0=not linked, 1=linked)

- Final groupings for TA were done on statistical significance to the reference group. (It has since been pointed out that we should have also considered sizes of parameter estimates when grouping territorial authorities.)
- Used results of regression to get an idea of relative importance of variables as well as ways the variables could be amalgamated.
- Final order of variables in decending order of importance was sex, age, TA, ethnicity, time since census, rurality, NZDep.
- Other than sex which was always kept separate, made 4 levels of age, TA, ethnicity, rurality, and 3 levels of time since census, NZDep. (With

the levels becoming grouped more i.e. Age2 has more levels than Age3)

- Needed to put ethnicity into prioritised ethnicity to give it just one value for each person.
- Each bias dataset was looked at separately to produce strata and starting level was decided for each variable, then "tweaks" were done for particular combinations that needed combining further. See Table 32 for main groups.

The SAS programme for this data management and analysis above can be found in Appendix 8 (CreateBiasWgtforCohortFinal.sas).

Table 32 Main Strata groupings for Bias Linkage weights

Bias Dataset	**Variable and level**	**Levels**
1981 and 1986	Age Level 2	0-24 yrs, 25-29 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 3	Statistical Significance groups : N.S., 0.10 Prob, 0.05 Prob, 0.01 Prob, Missing
	Prioritised Ethnicity Level 3	Maori, Pacific, NonMaoriNonPacific
	Time since census Level 3	Dates –ve&0-24 mths, 25-high mths
	Rurality Level 4	All Urban, NonUrban or Missing
	NZ Dep Level 3	Dec 1-6&Miss, Dec 7-10
1991, 1996 and 2001	Age Level 2	0-24 yrs, 25-29 yrs, 30-39 yrs, 40-49 yrs, 50-59 yrs, 60-69 yrs, 70-79 yrs, >=80 years
	Territorial Authority Level 3	Statistical Significance groups : N.S., 0.10 Prob, 0.05 Prob, 0.01 Prob, Missing
	Prioritised Ethnicity Level 2	Maori, Pacific, Asian, NonMaoriNonPacificNonAsian
	Time since census Level 3	Dates –ve&0-24 mths, 25-high mths
	Rurality Level 4	All Urban, NonUrban or Missing
	NZ Dep Level 3	Dec 1-6&Miss, Dec 7-10

Table 33 shows the summary of all people with cancer and those individuals linked to census records by link status, sex and cohort. There were many passes in the QualityStage™ linkage process and the Link Status column shows how good the links were. The status variable has been broken down into mesh block and area unit passes and whether the QualityStage™ linkage weights were high, medium or low. (Note for the QualityStage™ linkage process, high weights mean very good agreement on all variables and are better matches). It is noticeable that the majority of links were high QualityStage™ weights in the mesh block passes, followed by high weights in the area unit passes, showing that all of these are very good links and no variables were missing. Some of the linkages with lower QualityStage™ weights were because of missing variables on the cancer datasets, especially missing country of birth.

The following tables show that linkage of cancer records to census records was predicted by a number of socio-demographic factors.

- See above for explanation of high, medium, low weights (and they mean QualityStage matching/integration/linkage weights, not the linkage adjustment weights we added in datalab).
- Linkage success was highest if the cancer ocurred close to the time of census, and linkage success reduced the longer it was since the census, mainly due to people changing addresses. On the cancer files we tried to record as many meshblocks and area units, but in some cases would not have found the correct meshblock at census (possibly also due to different geocoding bases of the health system's domicile code (which relates to area unit) and the conversions we needed to do to them to make them base 2001). Linkage success also improved for the later cohorts compared to the earliest cohort (1981) where addresses had not been recorded as well and thus we could not geocode the address to 2001 mesh block and area unit codes. Instead we had to rely on the health data domicile code (equivalent to area unit but has different bases in different years and therefore needs to be forward coded to 2001 base).
- Time since census was split into two groups for bias linkage weighting: less than or equal to 24 months, greater than 24 months. Please note that 2001-02 cohort only has 46 months of follow-up, rather than 60 months because follow-up finished 31st December 2004.

- Linkage rates were lowest for 15 – 24 year olds both males and females, and 25 – 44 year old men. This reflects the usual patterns we see with linking to the census as these younger people are highly mobile and often do not have much interaction with the health system, hence it is more difficult to find their actual mesh block or area unit at the time of census, we only have these details for the time of the cancer. 25 – 44 year old females have a slightly better linkage rates because a large proportion of these may be involved in routine interactions with the health system concerning maternity.
- Until recently only one ethnicity was recorded on the various health systems and often it was assumed, not asked, and Asian ethnicity was not often recorded on earlier cohorts, whereas on the census people self-report their multiple ethnicities. This might be the reason for the lower linkage rates, especially for Asians in earlier cohorts in Table 36 (Total Ethnicity on the Cancer Registry) and Table 37 (Total Ethnicity on the NHI). We have included this NHI ethnicity table as the Cancer Registry will (or has already) stopped recording ethnicity and is instead using the NHI ethnicity for a person. In our table, NHI ethnicity linkage rates are just slightly better. There is a large increase in missing ethnicity on the cancer registry from 1991 cohort onwards, particularly in those 45-64 years.
- Linkage rates by Regional Health Authority (Table 38) are good and have been slowly increasing over the cohorts.
- There does not seem to be any differentialisation of linkage rates by rurality (Table 39). They all seem good rates.
- Similarly linkage rates are good for sime since census (Table 40) but rates do reduce slightly the further from the census date (and hence people may have moved from where they were at census).
- There do not seem to be any patterns with linkage rates for NZ Deprivation (Table 42)