



(Mis)classification of ethnicity on the New Zealand Cancer Registry: 1981–2004

Caroline Shaw, June Atkinson, Tony Blakely

Abstract

Background Māori and Pacific peoples are undercounted in cancer incidence statistics relative to census statistics. We use linked census and Cancer Registry data sets to determine the extent of misclassification between 1981 and 2004.

Methods The 1981, 86, 91, 96 and 2001 censuses were anonymously and probabilistically linked to individuals with a cancer registration for the entire intercensal period, or 31 December 2004 in the 2001 cohort. We compared counts by ethnicity between census and Cancer Registry data. Correction ratios and percentage under or overcounts are presented.

Results Undercounting of Māori and Pacific peoples was marked in the first cohort. For example Māori were undercounted 31% on the Cancer Registry compared to self reported ethnic origin in the 1981–86 cohort. Gradual improvements were seen and by the 2001 cohort undercounting was 15% and 10% for Māori and Pacific peoples respectively. For Asian people undercounting improved from 68% to 13% over the time studied. Reciprocally, non Māori/Pacific/Asian peoples were consistently overcounted.

Conclusion There is undercounting of Māori, Pacific and Asian events on cancer registration data, relative to census data, throughout 1981–2004. Cancer incidence rates need revising, to correctly understand the epidemiology and to inform cancer policy. Steps to improve the quality of ethnicity information remain a priority.

Cancer is now the single biggest cause of mortality in New Zealand.¹ Additionally it is a growing source of inequalities in health outcomes, in particular between ethnic groups.^{2–7} In recognition of the importance of cancer for population health there is increasing policy attention nationally (and internationally) on cancer control. Locally a Cancer Control Strategy and implementation plan has been developed, in order to provide a focus for action and subsequent monitoring.^{8–10}

In order to effectively plan, deliver and evaluate cancer control policy and interventions for their effect on inequalities accurate ethnicity statistics are essential. It is also a basic right of all ethnic groups to have accurate information on their health status.

Cancer mortality statistics are, currently, reasonably accurate by ethnicity and adjusters are available to correct for undercounting over the 1980s and 1990s.^{11 12} Cancer incidence statistics are calculated using numerator ethnicity information from the New Zealand Cancer Registry (NZCR) and denominator information from the census.¹³ In comparison to mortality statistics, the accuracy of incidence statistics by ethnicity is less clear. Table 1 shows the provenance (where it was possible to

establish) of information on ethnicity on the NZCR, and census datasets between 1981 and 2004.

Table 1 Origin of information on ethnicity on Census and Cancer Registry 1981–2004

Cancer Registry	Census
<p>1980–early 1994</p> <ul style="list-style-type: none"> • Cancer Registration Form sent to the Cancer Registry for each incident cancer from hospitals. Patient’s ‘race’ (Māori, Pacific Islander or Other) was recorded on the form. Unclear who filled out these forms. Assume that the information on patients race was obtained from hospital notes. • From early 1990s ethnicity taken from these National Minimum Dataset (NMDS; e.g. hospitalisation events) entries. (These NMDS events may have preceded the cancer event). 	<p>1981</p> <ul style="list-style-type: none"> • ‘biological’ ethnic origin • multiple groups allowed ancestry question
	<p>1986</p> <ul style="list-style-type: none"> • self-identified ethnic origin • multiple groups allowed
<p>1994–1999</p> <ul style="list-style-type: none"> • Reporting of incident cancers by laboratories on a Cancer Reporting Form. Detailed ethnicity options on this form, similar to census ethnicity question. Unclear who filled in the forms and where ethnicity information was obtained from (assume hospital patient management systems). • If no ethnicity was on the form, was obtained from the NMDS/National Health Index (NHI)/mortality collections. 	<p>1991</p> <ul style="list-style-type: none"> • self-identified ethnic origin • multiple groups allowed
<p>1999 onwards</p> <ul style="list-style-type: none"> • Incident cancers obtained from pathology labs sending forms to Cancer Registry. No ethnicity on these forms. Ethnicity obtained from NHI initially (again this may precede the cancer event). If no ethnicity available from NHI then staff updated it at a later date (from subsequent NHI, NMDS events or mortality collection). 	<p>1996</p> <ul style="list-style-type: none"> • self-identified ethnicity • more encouragement of multiple groups
	<p>2001</p> <ul style="list-style-type: none"> • self-identified ethnic origin • multiple groups allowed • same as 1991 question

Source: Personal communication Susan Hanna Ministry of Health May 2008, Statistics New Zealand (www.stats.govt.nz)

Inaccuracies in cancer statistics by ethnicity come from three sources. Firstly by definition there is numerator (i.e. NZCR data) denominator (i.e. census data) bias in the calculation of cancer incidence rates by ethnicity, as ethnicity information collection was not consistent between these two datasets over most of the time period studied. Secondly the ethnicity questions in both datasets have changed over this time which may alter the numbers in the numerator and denominator. Finally while there

has not (to our knowledge) been a formal audit of the accuracy of NZCR ethnicity information, one audit of lung cancer in Auckland and Northland showed that 3% of records in this region in 2004 were misclassified.¹⁴

Additionally audits of the accuracy of ethnicity information in other health datasets in the last three decades have (almost entirely) shown over counting of NZ European/other and undercounting of Māori and (when examined) Pacific People.^{15 16}
¹⁷ (The largely correct data of recent times for mortality data is an exception.¹²)

This paper describes uses linked census-cancer data from CancerTrends, a study that linked Cancer Registry and census records between 1981 and 2004, allowing us to compare self recorded ethnicity from census forms with ethnicity as it is recorded on the NZCR. This allowed us to calculate the extent of over and undercounting of different ethnic groups on the NZCR between 1981 and 2004 and provide adjustors for use by the health sector.

Methods

Study data—Five closed cohorts were created of the New Zealand usual resident population (all ages) on census night 1981, 1986, 1991, 1996, 2001, followed up for incident cancer(s) until the subsequent census or in the case of the 2001 cohort, until 31 December 2004 (the most recent data available at the time of the study). Cohorts were created using probabilistic record linkage software (QualityStage). The software linked anonymised census and Cancer Registry records within a geographic area (meshblock or census area unit) on sex, date of birth, ethnicity, and country of birth, using the same method as that in the New Zealand Census-Mortality Study (NZCMS).¹⁸⁻²¹

Table 2. Summary of data linkage results by cohort

Cohort	Usual resident population on census night	People with incident cancers in period of follow up	People with cancer who were linked to census record (%)	Positive predictive value (PPV) of links (%)
1981-86	3,143,307	52,699	73.2	95.2
1986-91	3,263,283	63,626	77.1	95.7
1991-96	3,373,926	77,159	79.2	95.1
1996-01	3,516,513	96,422	79.7	95.8
2001-04	3,630,534	83,789	81.7	96.9

Note: Each 5-year period is from census night to census night (about 7 March on average) except for 2001 cohort which ended on 31 December 2004. Note PPV can only be calculated on links made by linkage software passes, not on the proportion made through clerical review, and is therefore an estimate. Census counts were random rounded in accordance with Statistics New Zealand policy.

Table 2 shows the number of census records, individuals with incident cancer, the proportion of records linked and the positive predictive value of those links. The method for calculating PPV has been detailed elsewhere,¹⁸ and further detail of the record linkage is available elsewhere.²²

Ethnicity—A modified total ethnicity approach was used for this work. Total ethnicity places an individual in all ethnic groups that they identify with, thus capturing (most) multiple ethnic affiliations of individuals. Total ethnicity is the approach that Statistics New Zealand now recommends.²³ If individuals indicated any/all of Maori, Pacific and/or Asian ethnic affiliation they were placed in any/all of Total Māori, Total Pacific, Total Asian ethnic groups.

The residual people who did not indicate any of the above ethnic affiliations were placed in the non-Māori/Pacific/Asian (nMPA). The latter group is not a 'Total ethnicity' group, as strictly speaking in a true total ethnicity 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 we made this a residual category. Missing ethnicity is reported with nMPA in these results; however this is available separately if required.

The 1981 census question was based on ethnic origin rather than ethnic affiliation and blood quantum measures were used. In order to convert this into total ethnicity to be consistent with later years, we classified someone as Māori if they recorded any fraction as Māori, and likewise for Pacific and Asian.

Calculating the extent of misclassification of ethnicity on the cancer Registry—The methods used are a modification of, and improvement over, previous methods used in the NZCMS to be able to cope with 5 large cohorts of data.^{12 24}

It was initially determined what factors predicted linkage by identifying highly probable links (HPL). HPL are those census and cancer record that were linked without using ethnicity as a matching variable and comprise 61.6%, 67.6%, 71.2%, 75.7% and 69.5% for each consecutive cohort. These 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 (further details of the process will be published in future technical documents).

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).

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.

Approval was granted for this project under the Statistics New Zealand Data Integration Policy²⁵ and the Wellington Ethics Committee granted ethics approval for CancerTrends (Ref 04/10/093).

Results

The weighted ethnicity counts according to both cancer and census data, using a total definition of ethnicity, for all five cohorts (ages and sexes combined) are shown in Table 3. For example, in 1981–86, there were 1971 cancer registrants identified as Māori on the NZCR, but an estimated 2829 estimated as Māori according to census data. Ratios and percentage undercounts are both presented. The ratio, 1.44 (i.e. 2829/1971), denotes that the NZCR counts for Māori in 1981–86 need multiplying by 1.44 to give a ‘correct’ estimate of the ‘gold standard’ census Māori count. Expressed as a percentage undercount, there was a 30% undercount of Māori on the NZCR compared to the census (i.e. 1–1971/2829 or 1–1/1.44). It should be noted, however, that the 1981 census did not actually ask self-identified ethnicity, making strict time series comparisons difficult.

Table 3. Cross classified cancers, misclassification ratios and percentage under/overcounts for all cancers by ethnicity and cohort 1981–2004

Ethnicity	1981–86*				1986–91*				1991–96*				1996–01*				2001–04*			
	Census ethnicity	Registry ethnicity	Ratio [^]	Count [#] (%)	Census ethnicity	Registry ethnicity	Ratio [^]	Count [#] (%)	Census ethnicity	Registry ethnicity	Ratio [^]	Count [#] (%)	Census ethnicity	Registry ethnicity	Ratio [^]	Count [#] (%)	Census ethnicity	Registry ethnicity	Ratio [^]	Count [#] (%) [#]
Total Māori	2,829	1,971	1.44	-31	4077	3261	1.25	-20	5619	4473	1.26	-21	8526	6582	1.3	-23	6966	5925	1.18	-15
Total Māori	432	318	1.36	-26	936	765	1.22	-18	1122	915	1.22	-18	1995	1635	1.22	-18	1896	1713	1.11	-10
Total Asian nMPA [†]	48,228	50,400	0.96	4	57,666	59,466	0.97	3	69,600	70,230	0.99	1	83,808	78,636	1.07	-7	72,372	70,464	1.03	-3
Missing & nMPA	49,222	50,406	0.98	2	58,332	59,484	0.98	2	69,873	71,406	0.98	2	84,736	87,141	0.97	3	73,272	74,646	0.98	2

* 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 Registry [#] Percentage underestimation (if negative) or overestimation (if positive). Formula = (1 - 1/ratio)*100. 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.

Table 4. Misclassification ratios and percentage over/undercounts for all cancers by ethnicity, age group and cohort 1981–2004.

Ethnicity	1981–86*		1986–91*		1991–96*		1996–01*		2001–04*	
	Ratio [^]	Count (%) [#]	Ratio [^]	Count (%) [#]	Ratio [^]	Count (%) [#]	Ratio [^]	Count (%) [#]	Ratio [^]	Count (%) [#]
Total Māori										
0–14 yrs	1.49	-33	1.43	-30	1.3	-23	1.4	-29	1.23	-19
15–24 yrs	1.48	-32	1.26	-21	1.34	-25	1.41	-29	1.25	-20
25–44 yrs	1.45	-31	1.26	-21	1.27	-21	1.26	-21	1.17	-15
45–64 yrs	1.34	-25	1.21	-17	1.22	-18	1.27	-21	1.16	-14
≥65 yrs	1.62	-38	1.28	-22	1.26	-21	1.33	-25	1.17	-15
Total Pacific										
0–14 yrs			1.3	-23	1.27	-21	1.33	-25	1.24	-19
15–24 yrs	1.63	-39	2.44	-59	1.6	-38	1.71	-42	1.52	-34
25–44 yrs	1.36	-26	1.22	-18	1.2	-17	1.27	-21	1.18	-15
45–64 yrs	1.26	-21	1.06	-6	1.13	-12	1.13	-12	1.07	-7

Ethnicity	1981-86*		1986-91*		1991-96*		1996-01*		2001-04*	
	Ratio^	Count (%)#	Ratio^	Count (%)#	Ratio^	Count (%)#	Ratio^	Count (%)#	Ratio^	Count (%)#
≥65 yrs	1.4	-29	1.22	-18	1.22	-18	1.21	-17	1.01	-1
Total Asian										
0-14 yrs							0.85	18	1.31	-24
15-24 yrs					2.13	-53	2.4	-58	1.59	-37
25-44 yrs			4.04	-75	1.93	-48	1.45	-31	1.16	-14
45-64 yrs			2.75	-64	1.46	-32	1.32	-24	1.13	-12
≥65 yrs			3.3	-70	1.25	-20	1.16	-14	1.09	-8
nMPA†										
0-14 yrs	0.9	11	0.91	10	0.93	8	0.96	4	0.94	6
15-24 yrs	0.91	10	0.93	8	0.92	9	0.95	5	0.95	5
25-44 yrs	0.94	6	0.95	5	0.96	4	1.04	-4	1	0
45-64 yrs	0.97	3	0.97	3	1	0	1.12	-11	1.06	-6
≥65 yrs	0.96	4	0.98	2	1	0	1.05	-5	1.02	-2
nMPA† and Missing										
0-14 yrs	0.91	10	0.92	9	0.92	9	0.9	11	0.93	8
15-24 yrs	0.92	9	0.94	6	0.91	10	0.9	11	0.93	8
25-44 yrs	0.95	5	0.96	4	0.95	5	0.94	6	0.96	4
45-64 yrs	0.98	2	0.98	2	0.98	2	0.97	3	0.98	2
≥65 yrs	0.99	1	0.99	1	0.99	1	0.99	1	0.99	1

* 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 Registry # Percentage underestimation (if negative) or overestimation (if positive). Formula = (1- 1/ratio)*100. Results from cells with very small numbers have been suppressed.

Considering other ethnic groups and time periods, the following trends are evident. First, the Māori undercount was greatest in 1981–86 (31%), then between 20% to 23% in the late 1980s and entire 1990s, and dropped somewhat to 15% in 2001–04. Second, the undercount of Pacific peoples was consistently lower than that for Māori and also improved over time, with a 26% undercount initially, 18% throughout middle three cohorts, and 10% in 2001–04. Third, for Asian people there was a large undercount of 68% in 1986–91, but improving dramatically to 13%. Fourth, the larger size of the nMPA group buffers them against large over/undercounting; however with the rise of ‘missing ethnicity’ in the 1990s nMPA became slightly undercounted on the NZCR (3% in the 2001–04 cohort). However if the ‘missing ethnicity’ are included with nMPA ratios were just less than 1.

The number of records on the NZCR with missing ethnicity increased from negligible (presumably a default option was in place) in the 1980s to 2349 (1.5%) records in the 1991 cohort, to 17 004 (8.8%) in the 1996 cohort and then declined to 8361 (4.9%) in the 2001 cohort. Of these people only 399, 1290, 1389 respectively in each cohort also had ethnicity missing on their census form. Just over 90% of individuals with missing ethnicity on the NZCR were in the nMPA ethnic group on their census form.

Table 4 shows misclassification by age group. For Asian, Pacific peoples and Māori the misclassification appears to be worse in those under 24 age groups in the latter cohorts. For example in Pacific people in the 2001 cohort there was 7% undercounting of those age 45–64, whereas for the 0–14 and 15–24 age groups there was 19% and 34% undercount respectively.

Figure 1 shows differences in ratios by DHB for Māori. There is a tendency for the extent of misclassification to be worse in southern DHBs. The ratios by DHB for Pacific people and Asian ethnic group are too small to draw any conclusions and have not been included in this graph.

Figure 1 Misclassification ratios for Māori, both sexes by cohort and District Health Board 1981–2004

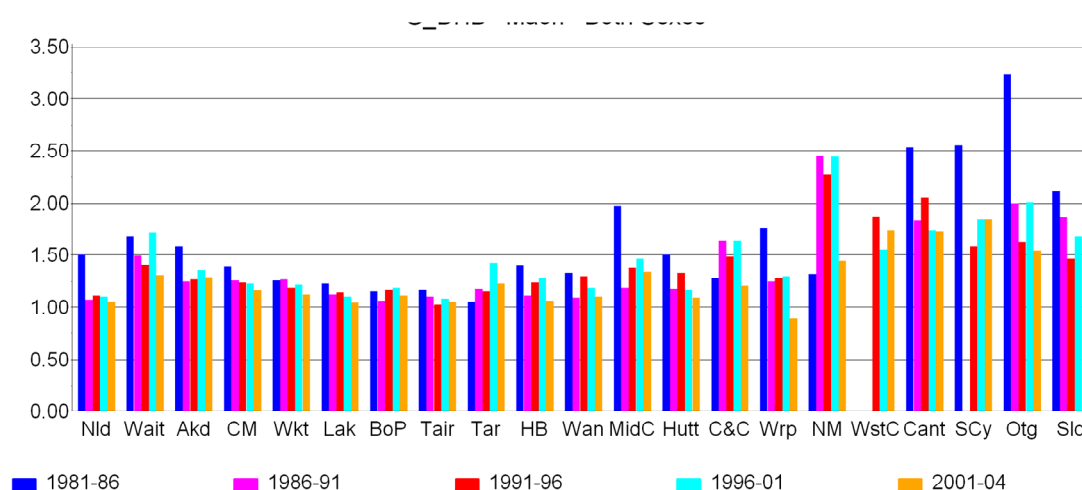


Table 5 shows misclassification ratios by cancer site. The table includes only a selection of the cancer specific ratios—others are available from the authors on request. These show that there is variation of misclassification by cancer site, for example in 2001 undercounting of Māori was greater for colorectal cancer (1.31) and melanoma (2.15) than overall (ratio 1.18) and less for lung cancer (1.06).

Table 5. Misclassification ratios for common cancer sites 1981–2004.

Sex	Cancer Site	Ethnicity	1981–86*	1986–91*	1991–96*	1996–99*	2001–2004*
Males and Females	Colorectal	Māori	1.46	1.36	1.31	1.48	1.31
		Total Pacific	1.41	1.23	1.39	1.31	1.12
		Total Asian		2.19	1.25	1.08	1.06
		nMPA+ & Missing	0.99	0.99	0.99	0.98	0.99
	Lung	Māori	1.21	1.23	1.14	1.09	1.06
		Total Pacific	1.23	1.1	1.17	1.04	1.02
		Total Asian		2.77	1.42	1.19	1.03
		nMPA+ & Missing	0.98	0.98	0.98	0.98	0.99
	Melanoma	Māori	4.57	1.74	2.44	3.48	2.15
		Total Pacific			2.09		
		Total Asian			2.2		2.63
		nMPA+ & Missing	0.98	0.99	0.99	0.98	0.99
Males	Prostate	Māori	1.46	1.36	1.38	1.64	1.28
		Total Pacific		0.89	1.19	1.26	1.09
		Total Asian			1.71	1.49	1.14
		nMPA+ & Missing	0.99	0.99	0.99	0.98	0.99
Females	Breast	Māori	1.61	1.28	1.36	1.31	1.16
		Total Pacific	1.62	1.14	1.23	1.22	0.98
		Total Asian		4.31	1.78	1.5	1.18
		nMPA+ & Missing	0.97	0.98	0.97	0.96	0.98

* 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 Registry. Results from cells with very small numbers have been suppressed.

Discussion

This paper presents adjustment ratios that can be used by the health sector to obtain corrected population estimates for incident cancer by ethnicity between 1981 and 2004. These ratios show that while there have been improvements over time in the accuracy of ethnicity data on the New Zealand Cancer Registry there remains a substantial undercount of Māori, Pacific and Asian ethnic groups in cancer registrations.

Strengths and limitations—The strengths of this study include having access to all cancer registrations between 1981 and 2004 and a gold standard ethnicity in the form of self identified ethnicity on the census form. This work is likely to have a high degree of accuracy due to high percentage and accuracy of records linked. In addition,

calculation of these ratios was done on a restricted dataset that excluded records that used ethnicity as a linking variable (which would have biased the results).

Limitations include that the ratios were not calculated on the entire dataset due to incomplete linkage as well as the need to restrict to the highly probable links; however methods were used to apply weights that have been shown to work well previously.^{20 26} Additionally the ethnicity questions in all but 2 censuses have been different so this is an imperfect time series. However, it is the best available and – most importantly – it is the actual denominator used in the calculation of rates. We also note that Pacific and Asian ethnic groups are not homogenous populations and within the broad misclassification ratios there may be differences by specific ethnic group i.e. Samoan, Indian.

Specific issues—The proportion of missing data is a marker of quality of Cancer Registrars.²⁷ The percentage of people on the Cancer Registry with missing ethnicity is reasonably substantial in later periods, e.g. 4.9% of records in the 2001 cohort. These are thought to be individuals who have cancer diagnosed in the private sector (personal communication Susan Hanna NZHIS May 2008). People with missing ethnicity are included in the non-Māori group for reporting purposes by NZHIS.¹³ The information in this study, showing that just over 90% of this missing group are nMPA, confirms that this is reasonable practice currently. However in other datasets this is not the case¹⁶ and ways to decrease the proportion of missing ethnicity data need to be explored.

Some of the differences in misclassification by cancer site may be due to chance and/or random variation due to small numbers; however for the larger cancer sites such as breast, lung and CRC it may reflect a real difference. Reasons for differences by cancer site are not obvious, potential suggestions could include age differences in cancer distribution, relationship to smoking (i.e. assumptions are made about ethnicity in the presence or absence of tobacco consumption). However this is an area that requires explanatory investigation.

Thinking about ethnicity on health datasets—This work shows that misclassification of ethnicity on the Cancer Registry has been present over the entire period studied, but the extent of misclassification varies by time, ethnic group, age, District Health Board (DHB) and cancer site. The ratios need to be applied to any historical time series in order to understand trends in cancer incidence for specific ethnic groups and to inform planning for cancer control activities. Alternatively, cancer incidence rates calculated directly from the linked census-cancer data (to be released in 2009) should be used. These corrected rates should also be used to provide accurate monitoring of the impact of the Cancer Control Strategy and Action Plan.^{8 9} DHBs should consider using the adjustment factors in this paper to ensure that funding is allocated based on an accurate picture of health need in their population.

Is there need for the ongoing linkage of cancer registrations and census data in order to calculate the level of misclassification of ethnicity? We would argue that there is for the following reasons.

There have been considerable efforts within the health sector in the last decades to improve the collection of ethnicity data, for example through promulgation of protocols around ethnicity data collection and associated training packages.²⁸ These

appear to have had mixed success. Classification of (total) ethnicity was largely accurate for mortality data between 2001–2004,¹² however for primary care collections for children there is evidence, at least in Waitemata DHB in 2005, of substantial misclassification of Māori and Pacific children.¹⁶

The reasons for the discrepant findings in the extent of misclassification of ethnicity in these different health collections remain obscure. The ethnicity question is theoretically standard throughout the health sector, using a close approximation of the census question.²⁸ In reality different actors and processes are involved. Mortality record ethnicity is mostly collected by funeral directors while ethnicity from the NZCR is (currently) largely taken from records obtained from hospital encounters, in which information is collected by different staff.

Ethnicity information in primary care collections may come from the NHI or may be self identified.¹⁶ While it could be hypothesised that the smaller number of funeral directors collecting information could allow a more standardised approach than primary, secondary or tertiary care there is no specific evidence to support this. Indeed there is some evidence to the contrary.²⁹ Moreover CancerTrends has shown that within the same health collection there is variation—there was more misclassification in the southern DHBs than northern DHBs.

The variability in misclassification between datasets does provide a good opportunity, however. There are datasets or parts of datasets that are largely ‘accurate’ and closer examination of the reason(s) for this (for example why have there been improvements in the mortality collection and what factors predict less misclassification in some DHBs) will potentially provide locally relevant evidence about what works to improve ethnicity data.

At a more systematic level it is worth reminding ourselves that inaccuracies in ethnicity data in health collections have been documented for at least 20 years in New Zealand.^{15,17} The persistent nature of this misclassification, and the consistent direction of misclassification, suggests that there may be implicitly accepted practices that permit the overcounting of the numerically dominant ethnic group at the expense of other ethnic groups.

Assuming that we cannot rely on ethnicity in health datasets to be directly comparable with census data for the calculation of rates, a number of options have developed to correct for misclassification. These are:

- Using the ever Māori approach (i.e. if a person has ever been recorded as Māori on any health event they are classified as Māori for the analysis).⁵
- Applying NZCMS mortality adjustors to non-mortality datasets.³⁰
- Calculating bespoke adjustors from available information.³¹

There are limitations with each of these methods, for example ever Māori approach now appears to over count Māori.³¹ The bespoke adjustors calculated for use on cancer incidence statistics in Hauora IV estimated that Māori were undercounted between 2–16% (depending on age group) on the NZCR between 2000 and 2005.⁴ Our work suggests that these figures are probably slightly conservative and the cancer incidence disparities described in that publication were slightly underestimated.

Finally, it is now apparent that mortality and cancer incidence adjusters are not interchangeable.³⁰

Hence we would argue that as misclassification in the Cancer Registry is present but variable by a number of factors, that misclassification is not consistent across health datasets, that the reasons for this misclassification are not understood and that other methods to correct it have not been entirely successful, therefore there is currently a need to continue linking these datasets.

It is apparent that getting accurate concordance between ethnicity data between health and census (usual denominator) datasets is difficult. Efforts to improve the quality of ethnicity data on health datasets must continue, and this study shows that further work in this area is needed. This could include further training, and instituting formal quality assurance and audits processes.

In addition to improving ethnicity data, we also propose another option for calculating rates for non-cancer, non-mortality health events. That is to consider using the National Health Index (NHI) file as the denominator in calculation of rates. This would be possible for all health datasets that can be linked to, or have data on ethnicity from, the NHI, and would be possible by a number of demographic factors such as sex, age, ethnicity, deprivation and region. This will completely obviate any numerator-denominator bias in the calculation of rates—for recent health events at least. However there are three main limitations to this approach:

- First, it will still not be possible to calculate accurate historical rates due to the NHI file being constantly updated and therefore only valid as a total population file in recent years.
- Second, whilst accurate for rates, actual numbers of event by ethnic group will still be biased when compared to census ethnicity (currently seen as the gold standard).
- Third, the NHI file probably includes many people who have migrated out of New Zealand, although this might be overcome by periodic linkage to either immigration data or PHO data.

We encourage the Ministry of Health, DHBs and health researchers to seriously consider and scrutinise the use of the NHI file for denominator data.

Conclusions

Although it has improved over time, misclassification of ethnicity remains a problem on the New Zealand Cancer Registry. Māori, Pacific and Asian peoples remain undercounted and non-Māori/Pacific/Asian are consequently over counted. Hence Māori, Pacific and Asian peoples have been under enumerated in cancer incidence statistics historically as well as contemporaneously. Such undercounting limits our ability to ensure that policy to improve cancer outcomes and reduce inequalities is targeted and evaluated appropriately.

Statistics New Zealand Security Statement: 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 authors, not Statistics New Zealand.

Competing interests: None known.

Author information: Caroline Shaw, Senior Research Fellow; June Atkinson, Senior Analyst; Tony Blakely, Research Professor; Department of Public Health, University of Otago, Wellington

Acknowledgments: CancerTrends is a Health Research Council funded study. This particular analysis was funded by the Ministry of Health (the views expressed in this paper do not necessarily reflect those of the Ministry of Health). The authors thank Statistics New Zealand for their ongoing support for CancerTrends as well as Ricci Harris for her thoughtful comments on an earlier draft of this paper.

Correspondence: Dr Caroline Shaw, Senior Research Fellow, Department of Public Health, University of Otago Wellington, PO Box 7343, Wellington South, New Zealand. Fax: +64 (0)4 3895319; email: caroline.shaw@otago.ac.nz

References:

1. New Zealand Health Information Service. Mortality and Demographic Data 2004. Wellington: Ministry of Health, 2007.
2. Fawcett J, Blakely T. Cancer is overtaking cardiovascular disease as the main driver of socioeconomic inequalities in mortality: New Zealand (1981–99). *J Epidemiol Community Health* 2007;61(1):59–66.
3. Blakely T, Ajwani S, Robson B, Tobias M, Bonne M. Decades of disparity: widening ethnic mortality gaps from 1980 to 1999. *New Zealand Medical Journal* 2004;117(1199).
4. Robson B, Harris R, editors. *Hauora: Maori Standards of Health IV*. Wellington: Te Ropu Rangahau Hauora a Eru Pomare, 2007.
5. Robson B, Purdie G, Cormack D. *Unequal Impact: Maori and Non-Maori Cancer Statistics 1996–2001*. Wellington: Ministry of Health, 2006.
6. Shaw C, Blakely T, Sarfati D, Fawcett J, Hill S. Varying evolution of the New Zealand lung cancer epidemic by ethnicity and socioeconomic position (1981–1999). [see comment]. *New Zealand Medical Journal* 2005;118(1213). <http://www.nzmj.com/journal/118-1213/1411>.
7. Shaw C, Blakely T, Sarfati D, Fawcett J, Peace J. Trends in colorectal cancer mortality by ethnicity and socio-economic position in New Zealand, 1981-99: one country, many stories. *Australian & New Zealand Journal of Public Health* 2006;30(1):64–70.
8. Minister of Health. *The New Zealand Cancer Control Strategy*. Wellington: Ministry of Health, 2003.
9. Ministry of Health. *The New Zealand Cancer Control Strategy Action Plan 2005 – 2010*. Wellington: Ministry of Health, 2005.
10. Cancer Control Council of New Zealand. *Mapping Progress: The First Two Years of the Cancer Control Strategy Action Plan 2005-2010* Wellington: Cancer Control Council of New Zealand, 2007.
11. Ajwani S, Blakely T, Robson B, et al. Unlocking the numerator-denominator bias III: adjustment ratios by ethnicity for 1981-1999 mortality data. *The New Zealand Census-Mortality Study*. *New Zealand Medical Journal* 2003;116(1175). <http://www.nzmj.com/journal/116-1175/456/>
12. Blakely T, Atkinson J, Fawcett J. Ethnic counts on mortality and census data (mostly) agree for 2001-04: *New Zealand Census-Mortality Study update* *New Zealand Medical Journal* 2008;121(1281). <http://www.nzmj.com/journal/121-1281/3242>

13. New Zealand Health Information Service. Cancer: New Registrations and Deaths 2004. Wellington: Ministry of Health, 2007.
14. Stevens W, Stevens G, Kolbe J, Cox B. Comparison of New Zealand Cancer Registry data with an independent lung cancer audit. *New Zealand Medical Journal* 2008;121(1276). <http://www.nzmj.com/journal/121-1276/3119>
15. Swan J, Lillis S, Simmons D. Investigating the accuracy of ethnicity data in New Zealand hospital records: still room for improvement. *New Zealand Medical Journal* 2007;119(1239). <http://www.nzmj.com/journal/119-1239/2103>
16. Bramley D, Latimer S. The accuracy of ethnicity data in primary care. *New Zealand Medical Journal* 2007;120(1264):U2779.
17. Harris R, Robson B, Reid P, Keefe V. CHE Cultural reviews: Ethnicity data collection- error rates and training effectiveness. Wellington: Te Ropu Rangahau Hauora a Eru Pomare, 1997.
18. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *Int. J. Epidemiol.* 2002;31(6):1246–52.
19. Blakely T, Salmond C, Woodward A. Anonymous linkage of New Zealand mortality and Census data. *Aust NZ J Public Health* 2000;24:92–5.
20. Fawcett J, Atkinson J, Herd P, Blakely T. Record linkage of census and mortality 2001-04 records: New Zealand Census-Mortality Study technical Report No. 6. Wellington: Department of Public health, University of Otago Wellington, 2008.
21. Hill S, Atkinson J, Blakely T. Anonymous Record Linkage of Census and Mortality Records: 1981, 1986, 1991, 1996 Census Cohorts. NZCMS Technical Report No. 3. Public Health Monograph Series. Wellington: Department of Public Health, Wellington School of Medicine and Health Sciences, 2002.
22. Lash B. Documentation of the process and results of linking cancer data with census data 1981 to 2001. Wellington: Statistics New Zealand, 2008.
23. Statistics New Zealand. Statistical Standard for Ethnicity 2005. Wellington: Statistics New Zealand, 2005.
24. Ajwani S, Blakely T, Robson B, Atkinson J, Fawcett J, Kiro C. Unlocking the numerator-denominator bias for the 1980s and 1990s. NZCMS Technical Report. Wellington: Department of Public Health, Wellington School of Medicine and Health Sciences, 2003.
25. Statistics New Zealand. Data Integration Policy. Wellington, 2006.
26. Fawcett J, Blakely T, Atkinson J. Weighting the 81, 86, 91 and 96 Census-Mortality Cohorts to Adjust for Linkage Bias. NZCMS Technical report No. 5. Public Health Monograph Series. Wellington: Department of Public Health, Wellington School of Medicine and Health Sciences, 2002.
27. Skeet R. Quality and quality control. In: Jensen O, Parkin D, editors. *Cancer Registration: Principles and Methods*. Lyon: International Agency for Research on Cancer, 1991:101-7.
28. Ministry of Health. Ethnicity Data Protocols for the Health and Disability Sector. Wellington: Ministry of Health, 2004.
29. Fifth year medical students public health project (Group 1). Determination of ethnicity through death registration. Unpublished report, 1998.
30. Curtis E, Wright C, Wall M. The epidemiology of breast cancer in Maori women in Aotearoa New Zealand: implications for ethnicity data analysis. *New Zealand Medical Journal* 2005;118(1209). <http://www.nzmj.com/journal/118-1209/1298>
31. Harris R, Purdie G, Robson B, Wright C, Zhang J, Baker M. Appendix 3: Estimating Maori hospitalisations and cancer registrations. In: Robson B, Harris R, editors. *Hauora: Maori Standards of Health IV*. Wellington: Te Ropu Rangahau Hauora a Eru Pomare, 2007.