Constructing whole of population cohorts for health and social research using the New Zealand Integrated Data Infrastructure

Jinfeng Zhao,[1] Sheree Gibb,[2] Rod Jackson,[1] Suneela Mehta[1] and Daniel J. Exeter[1]

Abstract

Objectives: To construct and compare a 2013 New Zealand population derived from Statistics New Zealand's Integrated Data Infrastructure (IDI) with the 2013 census population and a 2013 Health Service Utilisation population, and to ascertain the differences in cardiovascular disease prevalence estimates derived from the three cohorts.

Methods: We constructed three national populations through multiple linked administrative data sources in the IDI and compared the three cohorts by age, gender, ethnicity, area-level deprivation and District Health Board. We also estimated cardiovascular disease prevalence based on hospitalisations using each of the populations as denominators.

Results: The IDI population was the largest and most informative cohort. The percentage differences between the IDI and the other two populations were largest for males and for those aged 15-34 years. The percentage differences between the IDI and Census cohorts were largest for people living in the most deprived areas. The ethnic distribution varied across the three cohorts. Using the IDI population as a reference, the Health Service Utilisation population generally overestimated cardiovascular disease prevalence, while the Census population generally underestimated it.

Conclusions and implications: The New Zealand IDI population is the most comprehensive and appropriate national cohort for use in health and social research.

Key words: population denominator, Integrated Data Infrastructure, data linkage, health data

In New Zealand (NZ) and elsewhere, national census populations are commonly used by health and social researchers as denominator populations for estimating rates and proportions for a wide range of outcomes.[1,2] The census has often been considered a near-comprehensive source of information about individuals and households, and their demographic, social and economic characteristics. However, the NZ census is normally conducted every five years and only captures the population at a specific point in time. Moreover, the census, which captures NZ residents who respond on census night, suffers from population undercount issues. While the official national net undercount rate published by Statistics NZ in the Post-enumeration Survey for the 2013 Census was estimated at 2.4% (about 104,000 people), it was 2.6% for males, 4.8% for people aged 15-29 years, 6.1% and 4.8% for Māori and Pacific people respectively and 3.4% for the North Island excluding the Auckland and Wellington Regions.[3]

The official Estimated Resident Population (ERP) produced by Statistics NZ provides estimates for people who live in NZ at a given time, and accounts for the census undercount and people who were overseas temporarily at the time of the census.[4] The ERP is derived by adjusting the census usually resident population count for net census undercount, the estimated number of residents temporarily overseas on census night, natural population changes and net migration between census night and a given date.

While the ERP can be used as a population denominator or a reference to assess the effectiveness of other denominators, it is only an aggregated estimate, so it is not possible to directly link outcome data (i.e. numerators) to ERP denominators to form a tailored individual person-based dataset.[5] In addition, the accuracy of the ERP generally decreases as the time of interest gets further away from the census date.[6]

Health service utilisation (HSU, also known as health contact) populations are increasingly used as population denominators in NZ health research, given the ability to link to health outcomes using NZ's unique national health index number.[5,7-9] A HSU population includes all individuals who received any publicly-funded or subsidised health services or who enrolled with a primary health organisation (PHO) over a given time period. People are expected to enrol (or re-enrol) with a PHO every three years and in Quarter 3 2013, it was estimated that 96% of New Zealanders were enrolled.[5] HSU populations have three main advantages over census data for health-related analyses. First, they comprise individuals who utilised or had contact (i.e. through enrolment) with publicly funded services, which enables identification of health service performance gaps and population

1. Section of Epidemiology & Biostatistics, School of Population Health, The University of Auckland, New Zealand
2. Centre of Methods and Policy Application in the Social Sciences, The University of Auckland, New Zealand
Correspondence to: Dr Jinfeng Zhao, Section of Epidemiology & Biostatistics, School of Population Health, The University of Auckland, PB 92019, Auckland Mail Centre 1142, New Zealand; e-mail: jinfeng.zhao@auckland.ac.nz
Submitted: July 2017; Revision requested: November 2017; Accepted: January 2018
The authors have stated they have no conflict of interest.

Aust NZ J Public Health. 2018; Online; doi: 10.1111/1753-6405.12781

sub-groups that require improved access to health care.[5,8] Second, as a continuously updated, consistent and individually-linkable source of reported demographic information, they reduce the risk of numerator-denominator biases[10] that may occur when numerators and denominators are collected from different sources. Finally, selection bias (which affects more traditional but less complete sampling approaches) is reduced.[8] However, while HSU populations potentially capture more people than traditional sampling approaches, some selection bias still occurs through the exclusion of people who have not used publicly-funded health services over the given time period, either because they were healthy or chose not to seek care, or because they accessed private health care services in NZ or abroad.

Statistics NZ's Integrated Data Infrastructure (IDI) now enables multiple large national and regional longitudinal data sources to be systematically and securely linked at the individual-level, offering unprecedented opportunities for researchers and policy makers to obtain rich and comprehensive national datasets. The IDI includes a wide range of administrative and survey data comprising more than one terabyte of data at the present time and it is continually growing.[11] At the core of the IDI is a spine that aims to capture all individuals who have ever been residents of NZ. Currently the spine has more than nine million uniquely identified individuals derived from probabilistic linkages between three datasets, namely tax data from 1999 onwards, births data from 1920 onwards and visa data from 1997 onwards.[12] Datasets within the IDI are exactly (deterministically) matched where there are common unique identifiers available, or probabilistically linked where there are overlapping personal and demographic variables (e.g. name, sex and date of birth) at the individual level.[13] The ability to integrate datasets from multiple sources in one environment opens up many new possibilities.

The aims of this research were to develop more complete and accurate, multi-source national population denominators from the IDI for health and social research; to demonstrate the extent to which IDI-derived population denominators are more complete than existing national population denominators; and to compare the performance of an IDI population denominator, an HSU population denominator, and a Census population denominator for estimating the prevalence of cardiovascular disease (CVD) based on hospitalisations.

Methods

Constructing an IDI usual resident population cohort

The study covers a one-year period from 6 March 2012 to census day on 5 March 2013. We slightly modified an "activity-based" approach developed by Gibb et al.[6] to construct a population cohort using datasets available within the IDI, which are records of an individual's interaction (i.e. activity) with a particular centrally-funded service (e.g. education, taxation, health, immigration). These records of activity can be used to determine whether individuals were present and resident in NZ for at least six months during study period. Individuals who met all of the following criteria were selected and linked to form the IDI population cohort:

- were within the IDI Spine; and
- were active in at least one of the following data sources: health, tax, education, and injury claims in the in-scope year, or in the births or visa dataset in last five years; and
- lived in NZ for more than six months of the in-scope year; and
- were alive on census day 2013; and
- were aged between 0 to 115 years.

In general, activity in the health, tax, education and injury datasets indicates that an individual was present in NZ in the in-scope year. The last five years of the births and the visa datasets enabled identification of children aged under five years of age, who may not have had an activity record in the other four datasets. The purpose of the visa dataset is to capture children under five years of age who immigrated to NZ from overseas. In theory, these six datasets together should provide good coverage over the whole age range.[6] We then used border movements data to exclude people from the population if they were only present in NZ for a short time (e.g. visitors using health services).[6]

Constructing a HSU cohort

The HSU population was created solely using Ministry of Health datasets that are available in the IDI using a similar definition to Chan et al.[5,8] Since the PHO collection is a quarterly return, we defined our cohort as at 31 March 2013 – the end of the closest quarter to Census night. Individuals who met the following criteria were selected and linked to form the HSU cohort:

1. were active in at least one of the following datasets: general medical subsidy claims, community laboratory test claims, the national non-admitted patient collection, the community pharmaceutical collection, and the publicly-funded hospitalisation discharge collection, in the period from 1 April 2012 to 31 March 2013; or
2. were enrolled in a PHO and met either of the following criteria:
   - were included in the PHO enrolment data submitted between the 2nd quarter of 2012 and the 1st quarter of 2013 inclusive; or
   - were included in the PHO enrolment data submitted in the 2nd or 3rd quarter of 2013 and whose last consultation dates or enrolment dates were in the period from 1 April 2010 to 31 March 2013; and
3. were alive on 31 March 2013; and
4. had NZ resident status as indicated in the population cohort demographics dataset; and
5. were aged between 0 to 115 years.

Constructing the 2013 census usually resident population

The census usual resident (hereafter referred to as the 'Census') population was solely extracted from census datasets that are available in the IDI. Individuals were included in the Census population if they were living in NZ on census day in 2013 and were identified as NZ adults or NZ Children in the census individual dataset.

Linking geographical and deprivation information to the three cohorts

All datasets include, or can be linked to, meshblock-level residential address information. Meshblocks are the smallest census geographical area in NZ (mean 2013 Census population=91). The date of the recorded meshblock information was the most recent available date using the end of the study period as a cut-off point. Individuals were assigned their meshblock by linking the following datasets: the address notification full table for the IDI cohort; the population cohort NHI address dataset and, if needed, the population cohort PHO address dataset, for the HSU cohort; and the census 2013 address dataset for the Census cohort.

A meshblock concordance table was then linked to the three populations with current meshblock as the key. Other area level information, such as the NZ Index of Deprivation 2013 (NZDep2013) and District Health Board (DHB), can then be linked using corresponding meshblock (in this case 2013 meshblock) in the IDI.

Extracting demographic information for the three cohorts

For the HSU and Census populations, age, sex and ethnicity information was sourced solely from the health population cohort demographics dataset and the census individual dataset respectively. It is important to note that these different data sources record ethnicity information differently, which will account in part for the differences observed between the cohorts. For example, the percentage of Māori in the HSU population (13%) was less than the percentage of Māori in the Census and IDI populations (14% and 16% respectively).

For the IDI population, the ethnicity information was sourced preferentially from the census data if available, or otherwise from the health data. If unavailable in either of those two sources, ethnicity was determined from the personal details dataset, which records Statistics NZ's best estimate of demographic information derived from multiple collections in the IDI using a set of specific rules. Age and sex information was sourced from the personal details dataset.

A prioritised ethnicity[14] variable, which assigns one ethnicity to each individual, was created for all three populations using the following prioritisation rules: An individual's ethnicity was defined as: (1) Māori if any of the person's ethnic codes was Māori. (2) Pacific if any of the person's ethnic codes was Pacific and none of them was Māori. (3) Asian if any of the person's ethnic codes was Asian and none of them were Māori or Pacific. Asian individuals were further defined into three prioritised sub-groups in the following order: (a) Indian; (b) Chinese; (c) Other Asian. If individuals had both Pacific and Indian codes, they were assumed to be Fijian Indian and were classified in the Indian category. (4) All Other ethnicities if none of the above apply.

Estimating CVD prevalence in the three populations

Individuals were identified as having a history of CVD if they: (1) had at least one hospitalisation for atherosclerotic CVD or haemorrhagic stroke between 1 January 1988 and census day 2013 recorded in the Publicly Funded Hospital Discharge Events and Diagnosis datasets; or, (2) were dispensed anti-anginal medications on at least three occasions in the five years preceding census day 2013 recorded in the Pharmaceutical database. The history of CVD information extracted at the individual level from health datasets was then linked to corresponding individuals within the three population denominators to calculate CVD prevalence by various demographic sub-groups. Note that individuals identified as having a history of CVD were present in all denominators.

Comparing the three populations and their associated CVD prevalence

Population counts and percentage differences were compared by gender, age, ethnicity, NZDep 2013 and DHB among the three populations.

To demonstrate the influence of different population denominators, CVD prevalence was compared by age, gender and ethnicity for people aged 25 years and above among the three populations. Percentage differences were calculated using the IDI population as the reference and a positive percentage means that the IDI population was greater than the HSU or Census populations.

All analyses were undertaken using SAS Enterprise version 7.1. The SAS codes developed can be accessed at http://nihiviewprd01.its.auckland.ac.nz/IDIPaper_additional_material/create_NZ_population_cohorts_in_the_IDI.sas.

Ethics approval

Ethical approval for this study was first granted by the Multi-Region Ethics Committee in 2011 (ref: MEC/11/EXP/078) with subsequent approvals from the Health and Disabilities Ethics Committee.

Results

Demographic comparison between the IDI, HSU and Census populations

There were 4,414,287 people included in the IDI population cohort as at census day in 2013. The HSU (4,266,789) and Census (4,242,051) populations included 3% and 4% fewer people respectively (Table 1).

The IDI population had 21,713 (-0.5%) fewer people than the 2013 ERP (4,436,000) but as the ERP is an estimated population and cannot be linked to other information at the individual level, only the IDI, HSU and Census populations were compared in the following sections (Table 1 and Figure 1).

Comparing distributions by gender

The IDI population included both more males and females than the HSU and Census populations. The difference was greater for males: 5% between IDI and HSU for males and 2% for females, and 5% between IDI and Census for males and 3% for females.

HSU versus IDI distributions for Asian males and females were somewhat asymmetrical. Males aged 19-30 years in the HSU population had much lower counts than in the IDI and Census populations, and while females aged 17-28 years in the HSU population also had lower counts, the difference was considerably less.

Comparing distributions by age and ethnic groups

In general, the percentage differences between the IDI, HSU and Census populations were largest for people aged 15-34, and were generally much smaller for people aged 55-64 years. Ethnicity data were unavailable for 230,646 people (5%) in the Census population.

The percentage differences by ethnicity between the IDI and the HSU and Census populations were large and highly variable for Māori, Pacific and Asian people aged 75-80 years and above due to relatively small population numbers in these groups. Therefore, these age groups are not compared in the description below.

The percentage difference between the IDI and HSU populations by ethnicity was greatest among Asian people (27%, particularly for Asian males aged 20-30 years), followed by Māori (21%), Pacific (5%) and All Other (-5%) ethnic groups. In the comparison between the IDI and Census populations, Māori and Pacific people had the largest percentage difference (both at 15%), while Asian people had the smallest difference (5%). However, the IDI-Census differences among Asian sub-groups varied dramatically, ranging from 11% among Other Asian individuals to -2% among Chinese people (i.e. the IDI population was smaller than the Census population for Chinese people).

Generally the IDI population for Māori and Asian people by age was larger than the Census and the HSU equivalents. However, the IDI population was smaller than the

Table 1: The IDI population compared to the HSU and 2013 Census populations by demographics, deprivation and DHB.

Distribution	Populations						Percentage Difference	
	IDI (N)	IDI (%)	HSU (N)	HSU (%)	Census (N)	Census (%)	IDI-HSU (%)	IDI-Census (%)
Total	4,414,287	100	4,266,789	100	4,242,051	100	3	4
Sex								
Male	2,172,804	49	2,063,709	48	2,064,018	49	5	5
Female	2,241,483	51	2,202,120	52	2,178,030	51	2	3
Unknown	6		960					
Age								
14 and under	896,934	20	877,863	21	865,629	20	2	3
15-24	635,931	14	592,158	14	586,446	14	7	8
25-34	556,155	13	523,173	12	514,686	12	6	7
35-44	588,114	13	573,150	13	573,273	14	3	3
45-54	615,594	14	607,929	14	601,629	14	1	2
55-64	494,421	11	494,805	12	493,350	12	0	0
65-74	357,531	8	341,604	8	346,134	8	4	3
75-84	194,913	4	184,020	4	187,584	4	6	4
85 and over	74,697	2	72,093	2	73,314	2	3	2
Prioritised ethnicity								
Māori	700,941	16	551,895	13	598,602	14	21	15
Pacific	288,249	7	274,005	6	244,158	6	5	15
Asian	482,670	11	354,516	8	457,167	11	27	5
Indian	163,287	4	113,412	3	151,809	4	31	7
Chinese	157,872	4	120,465	3	161,769	4	24	-2
Other Asian	161,508	4	120,642	3	143,589	3	25	11
All Other	2,942,430	67	3,086,376	72	2,711,472	64	-5	8
Unknown					230,646	5		
NZDep 2013								
Decile 1	437,673	10	408,354	10	412,083	10	7	6
Decile 2	445,056	10	409,140	10	414,828	10	8	7
Decile 3	437,592	10	397,764	9	404,820	10	9	7
Decile 4	428,610	10	389,430	9	392,961	9	9	8
Decile 5	428,769	10	386,922	9	388,773	9	10	9
Decile 6	427,437	10	385,911	9	383,445	9	10	10
Decile 7	427,929	10	386,034	9	377,427	9	10	12
Decile 8	435,309	10	392,229	9	375,891	9	10	14
Decile 9	446,430	10	404,058	9	373,101	9	9	16
Decile 10	458,928	10	425,415	10	360,126	8	7	22
Unknown	40,551	1	281,535	7	358,596	8		
DHB (ordered from north to south)								
Northland	159,291	4	148,269	3	130,515	3	7	18
Waitemata	540,732	12	474,141	11	489,000	12	12	10
Auckland	453,540	10	385,242	9	399,288	9	15	12
Counties Manukau	498,246	11	443,661	10	429,822	10	11	14
Waikato	371,061	8	350,178	8	330,573	8	6	11
Lakes	104,349	2	99,975	2	88,404	2	4	15
Bay of Plenty	215,970	5	191,991	4	186,852	4	11	13
Tairawhiti	46,737	1	46,173	1	38,442	1	1	18
Taranaki	113,697	3	108,114	3	101,955	2	5	10
Hawke's Bay	158,190	4	149,259	3	138,078	3	6	13
Whanganui	61,260	1	49,011	1	53,967	1	20	12
MidCentral	166,749	4	153,336	4	149,646	4	8	10
Hutt Valley	142,818	3	141,030	3	128,547	3	1	10
Capital and Coast	289,077	7	266,202	6	262,569	6	8	9
Wairarapa	41,793	1	38,484	1	38,043	1	8	9
Nelson Marlborough	140,391	3	134,868	3	126,678	3	4	10
West Coast	31,503	1	28,968	1	27,504	1	8	13
Canterbury	491,421	11	457,323	11	448,365	11	7	9
South Canterbury	56,712	1	53,571	1	52,017	1	6	8
Southern	304,365	7	277,884	7	274,596	6	9	10
Unknown/Outside DHB	26,388	1	269,106	6	347,199	8		

HSU population for Pacific children aged 3-13 years, Pacific females aged 27-35 years, All Other people aged < 64 years, and particularly for All Other people in the 15-35 years age group.

Comparing distributions by geographic level information

Comparing distributions by NZDep 2013 at the meshblock level

The IDI included more people than the HSU and Census populations in all NZDep 2013 deciles. The percentage difference between the IDI and the Census populations gradually increased as NZDep 2013 deciles increased (from 6% at decile 1 to 22% at decile 10). In contrast, the percentage difference in area deprivation between the IDI and the HSU populations was more consistent across deciles (7-10%). Fewer than 1% (40,551) of the IDI population did not have a matching NZDep 2013 decile. In contrast, more than 6% (281,535) and 8% (358,596) of the HSU and Census populations respectively were missing NZDep 2013 information at the meshblock level (Table 1).

Comparing distributions by DHB

The IDI included more people than the HSU and Census populations in all DHBs. The greatest percentage differences between the IDI and HSU populations occurred in the Auckland and Whanganui DHBs. In contrast, the greatest percentage differences between the IDI and Census populations occurred in the Northland, Tairawhiti, and Lakes DHBs where the proportion of Māori people was relatively high. The percentage differences between the IDI and HSU populations were much more variable between DHBs than the percentage differences between the IDI and Census populations. DHBs in the south of the North Island and throughout the South Island tended to have the lowest percentage differences between the IDI, HSU and Census populations. A population cartogram showing the percentage differences between the IDI and the HSU and Census populations by DHB can be viewed at http://nihiviewprd01.its.auckland.ac.nz/IDIPaper_additional_material/population_cartogram.png.

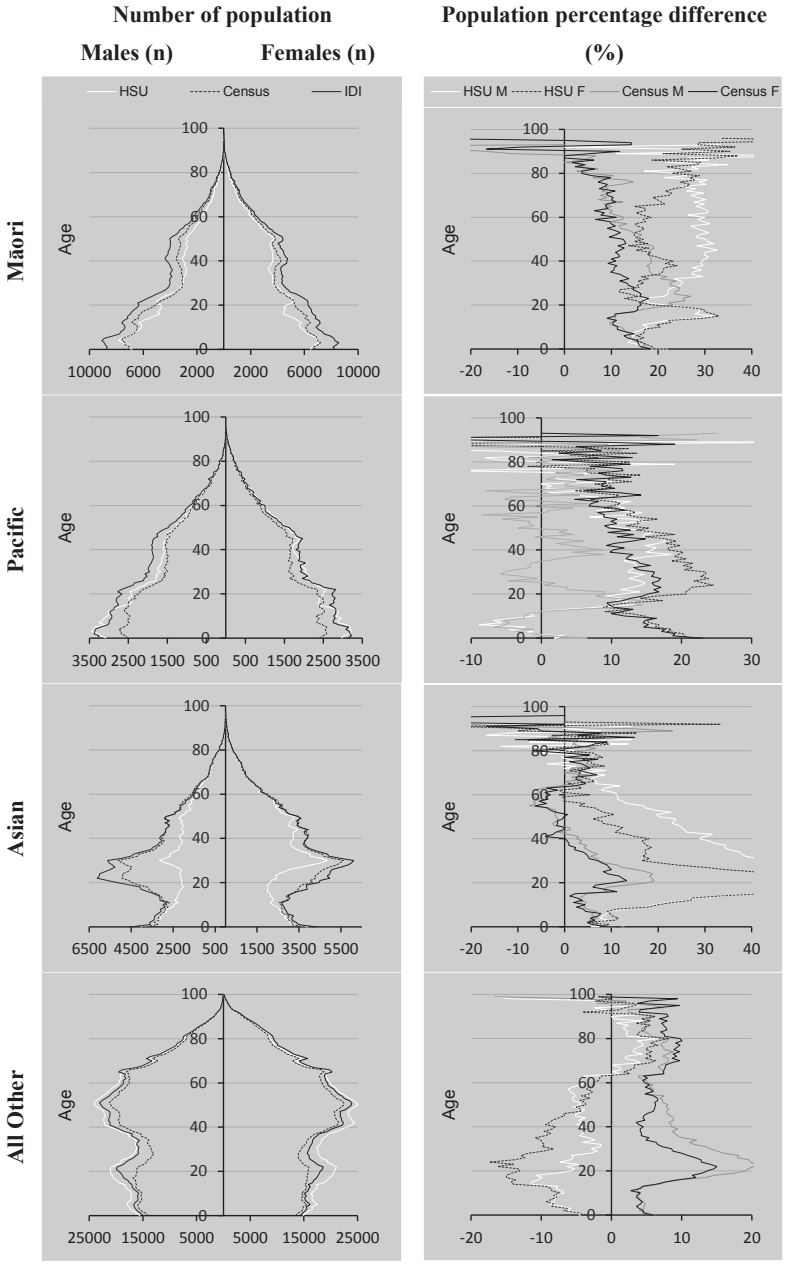
Comparing CVD prevalence in the IDI, HSU and Census populations aged >25 years

CVD prevalence among people aged 25 years or older increased with age across all ethnic and gender groups (Figure 2 and Supplementary Table) for all three

populations (IDI n=2,881,431, HSU n=2,796,288 and Census n=2,641,134). Males had consistently higher rates than females.

Using CVD prevalence from the IDI population as a reference, the HSU population generally overestimated CVD prevalence, except for Indian people, while the Census population generally underestimated it, except for the All Other people category. The degree of the discrepancy varied with age, gender and ethnicity, although there were some exceptions. For example, in the HSU population, CVD prevalence was overestimated to a greater degree for Māori and Pacific patients aged 65 years and above, especially for Māori men aged 75-84 years and Māori women aged 85 years and above. On the other hand, CVD prevalence was underestimated for Indian men aged 65 years and above and Indian women aged 65-84 years. In the Census population, CVD prevalence was overestimated for All Other men aged 75-84 years and All Other women aged 75 years and above, but was underestimated for Indian people aged 75 years and above.

Figure 1: Comparison between three NZ populations: Counts and percentage differences by age, gender and ethnicity.



Discussion and Conclusions

We used the Statistics NZ IDI environment to create a national individual-level cohort called the IDI population and compared this population with HSU and Census populations for the same period. Overall the total IDI population count was close to the official ERP estimate but included more individuals than the HSU or Census populations.

Using the IDI population as a reference, we made several observations. First, the percentage differences between the IDI, HSU and Census populations were larger for males, for those aged 15-34 years, and for areas located north of the MidCentral DHB. This corresponds to Statistics NZ findings regarding net census undercounts.[3] Second, more than 99% of the IDI population was linked to NZDep 2013 information at the meshblock level using multiple data sources available in the IDI. Where NZDep 2013 information was missing, this was partly because 0.5% (221) of meshblocks were not assigned a NZDep 2013 value originally.[15] However, more than 6% of the HSU population and 8% of the Census population was without NZDep 2013 information, mainly because of missing meshblock information in these datasets. Missing meshblocks in the Census population are largely a result of illegible or incomplete address information and out-of-scope (e.g. overseas) addresses.[16] Third, the percentage difference between the IDI and HSU populations was largest among young and middle-aged Asian men and Māori males, which suggests that these groups engaged less with NZ health services than their counterparts from other ethnic groups. For Māori men, known barriers to utilising health services include the high cost of healthcare, unavailability of services, accessibility issues and negative healthcare experiences.[17,18] In addition, the attitudes, beliefs and preferences of Māori male patients may not be understood by many healthcare providers.[18] For Asian peoples, reduced utilisation of health services could be related to a lack of knowledge among recent immigrants regarding how NZ health services are organised and/or utilisation of alternative medications such as natural therapies. By contrast, for All Other people aged 65 and below, the HSU population counts were greater than in the IDI cohort which could in part be due to European New Zealanders who live mainly overseas, returning to NZ for healthcare. For Māori males aged 40 years and above in the HSU population, the large percentage difference (about 30% fewer than in the IDI population) may lead to bias in estimating the burden of health outcomes.

This was evidenced by higher CVD prevalence for this group using the HSU rather than the IDI denominator. Finally, the percentage differences between the IDI and Census populations were greater for Māori and Pacific people (both 15%), for people living in areas of greater socio-economic disadvantage (22% greater for NZDep 2013 Decile 10) and for males aged 20-40 years. This indicates that census coverage was poorer for these groups, which will lead to biased results in health research when using the Census population as a denominator without appropriate adjustment. In addition, ethnicity information was missing for more than 5% of the Census population. The composition of this 5% population should be further investigated.

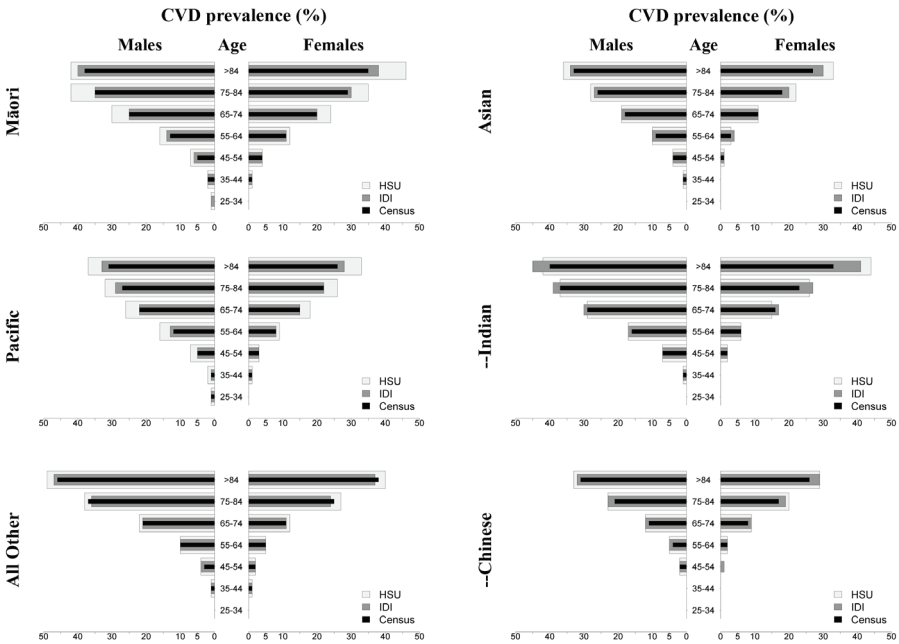
We identified large differences in distributions by ethnicity between the three denominator populations. While this is likely to be largely due to differential health service utilisation and census completion rates, it is also relevant that the three populations draw their ethnicity information from different data sources. In addition, there is known misclassification of ethnicity in health datasets prior to mid-2013, but this is likely to be reduced in more recent health data.[19] We acknowledge that a prioritised classification of ethnicity may add more complexity in unlinked numerator and denominator datasets. Boyd et al.[20] recommend using 'total' ethnicity counts (e.g. an individual's ethnicity is assigned as Māori if any ethnic group identified is Māori) when using unlinked census data as the denominator in the calculation of health outcome rates to avoid numerator-denominator bias.

Despite both the Census and HSU denominators having fewer people than the IDI denominator, CVD prevalence was generally overestimated when using the HSU but underestimated using the Census denominators. The reason is that more people with CVD were missing from the Census numerator than the HSU numerator because they did not complete the census. The highest overestimate of CVD prevalence using the HSU denominator was for Māori people, due to the high percentage difference between the HSU and IDI populations for this sub-group (Figure 1). Exceptionally, CVD prevalence was overestimated for All Other people aged 75 years and above using the Census denominator.

The differences in CVD prevalence by ethnicity between the three cohorts are due to a combination of selection and information biases, which vary between the cohorts. While numerator-denominator bias (a selection bias) will account for some of the differences observed, differential misclassification of ethnicity is also important.

In this paper we were primarily interested in the potential for numerator-denominator bias, i.e. bias effecting the comparisons of counts and we have used measures of CVD prevalence to illustrate the impact of this bias. However, we acknowledge that the use of more comprehensive numerator and denominator data does not address other forms of selection bias relevant to aetiological analyses.

Figure 2: CVD prevalence by age, gender and ethnicity using the IDI, HSU and Census populations as denominators.



Strengths of the IDI population cohort

First, the IDI population includes individuals who are excluded from the other two populations and therefore provides the most comprehensive denominator for population-level analyses. Individuals who were not in NZ on census night or did not fill out a census form are not captured in the Census population and the HSU will miss individuals who have not had contact with publicly funded health services in the given time period. Second, the IDI improves and enriches individual-level information about populations such as their ethnicity and incomes using cross-sourcing and prioritisation techniques. Third, the IDI increases the comprehensiveness of area level geographical information (99%), which varies markedly in individual datasets, e.g. 48% in the Ministry of Education dataset.[21] This creates opportunities to analyse health and social data at the individual level with associated geographical, environmental and contextual factors. It also paves the way to model individuals nested in neighbourhoods. In addition, analysing data at the meshblock level minimises the risk of confounding when analysing aggregated data.

Using an IDI population also reduces numerator-denominator biases since individuals in numerators, such as people with CVD, were included in population denominators. Finally, the IDI population can be constructed continuously over time because it was based on routinely collected administrative datasets that are updated regularly.

Limitations of an IDI population cohort

The IDI uses a probabilistic linking process to combine data, which inevitably produces linkage error. However the false positive linkage rate (i.e. if records for two different individuals are linked when they should not be) was estimated at less than 1%.[6] Linkage error can cause IDI population over-coverage or under-coverage, but is likely to be minimal. Another limitation is that Māori researchers have raised the issue of indigenous data sovereignty with regards to the IDI, and this might be a key area of focus in future discussions.[22]

In this study, we applied a one-year activity period, which led to some population under-

coverage. For example, residents who were not active in health, tax, education, and injury claims datasets in the year to 5 March 2013 were missed. Widening the activity period to a two- or three-year period could potentially reduce this undercoverage.[23]

Also, some non-residents might be included if they stayed in NZ for more than six months in the in-scope year and were active in the above datasets. The use of immigration datasets to identify individuals' immigration status could potentially reduce this over-coverage.

Implications

The IDI environment offers significant opportunities for health and social research. It promotes the use of existing administrative datasets, offers comprehensive information, is cost effective and avoids duplication of effort. The IDI also offers researchers valuable flexibility in extracting, revising and restructuring information as their research evolves.

It is exciting that the comparisons we show here provide insights into the structure of, and differences in, the three population cohorts and offer explanations for discrepancies in CVD prevalence observed using different denominator populations.

This research has constructed the most comprehensive and complete population cohort possible for NZ to date. It covers the entire country, and includes linked personal information and associated small area level information suitable for health and social research. The cohort can be readily linked to health and social economic variables at the individual level over time and space, opening up important research opportunities. Accurate denominator estimates are especially important for small population sub-groups, where even small discrepancies in the precision of denominator estimates may affect the magnitude of the outcome. Furthermore, accurate estimates of disease prevalence and disease outcomes using the most comprehensive population cohort available facilitate robust health service planning that can appropriately meet the needs of population sub-groups and improve health equity. However, constructing population cohorts from the IDI is a new field of research and ongoing work is needed (e.g. to understand the best ways to use ethnicity information from the IDI) to improve its validity and completeness.

Acknowledgements

This research is funded by the Health Research Council of NZ and the Virtual Health Information Network (NZ). We would like to thank Dr Katrina Poppe and Billy Wu for their advice on CVD data and related definitions, and Ying Huang for identifying a method for creating the charts in Figure 1. We thank Dr Matire Harwood for her valuable comments on issues related to Māori. We express our gratitude to the anonymous reviewers for their helpful advice. We thank the IDI team at Statistics NZ for their input and use of data.

Statistics New Zealand Disclaimer

The results in this paper are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Statistics NZ.

The opinions, findings, recommendations, and conclusions expressed in this paper are those of the author(s), not Statistics NZ.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation and the results in this paper have been confidentialised to protect these groups from identification.

Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy Impact Assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

References

1. Disney G, Teng A, Atkinson J, Wilson N, Blakely T. Changing ethnic inequalities in mortality in New Zealand over 30 years: Linked cohort studies with 68.9 million person-years of follow-up. *Popul Health Metr*. 2017;15(1):15.
2. Woodward A, Blakely T. *Healthy Country? A History of Life and Death in New Zealand*. Auckland (NZ): Auckland University Press; 2015.
3. Statistics New Zealand. *Coverage in the 2013 Census Based on the New Zealand 2013 Post-enumeration Survey* [Internet]. Wellington (NZ): Government of New Zealand; 2014 [cited 2017 Jan 20]. Available from: http://www.stats.govt.nz
4. Health Partners Consulting Group. *Southern District Health Board Health Profile*. Auckland (NZ): HPCG; 2014.
5. Chan WC, Papaconstantinou D, Winnard D. Service planning implications of estimating Primary Health Organisation enrolment rate based on a Health Service Utilisation population rather than a Census-derived population. *N Z Med J*. 2015;128(1418):52-64.
6. Gibb S, Bycroft C, Matheson-Dunning N. *Identifying the New Zealand Resident Population in the Integrated Data Infrastructure (IDI)* [Internet]. Wellington (NZ): Government of New Zealand; 2016 [cited 2017 Jan 18]. Available from: http://www.stats.govt.nz
7. Ministry of Health and Accident Compensation Corporation. *Injury-related Health Loss: A Report from the New Zealand Burden of Diseases, Injuries and Risk Factors Study 2006–2016* [Internet]. Wellington (NZ): Government of New Zealand; 2013 [cited 2017 Feb 9]. Available from: http://www.health.govt.nz and http://www.acc.co.nz
8. Chan WC, Jackson G, Wright CS, Orr-Walker B, Drury PL, Boswell DR, et al. The future of population registers: Linking routine health datasets to assess a population's current glycaemic status for quality improvement. *BMJ Open*. 2014;4(4). doi:10.1136/bmjopen-2013-003975
9. Telfar Barnard LF, Baker MG, Hales S, Howden-Chapman P. Novel use of three administrative datasets to establish a cohort for environmental health research. *BMC Public Health*. 2015;15(1):246.
10. Blakely T, Robson B, Atkinson J, Sporle A, Kiro C. Unlocking the numerator-denominator bias. I: Adjustments ratios by ethnicity for 1991-94 mortality data. The New Zealand Census-Mortality Study. *N Z Med J*. 2002;115(1147):39-43.
11. Statistics New Zealand. *Integrated Data Infrastructure* [Internet]. Wellington (NZ): Government of New Zealand; 2016 [cited 2016 Aug 1]. Available from: http://www.stats.govt.nz
12. Black A. *The IDI Prototype Spine's Creation and Coverage* [Internet]. Statistics New Zealand Working Paper No.: 16–03. Wellington (NZ): Government of New Zealand; 2016 [cited 2017 Feb 18]. Available from: http://www.stats.govt.nz
13. Statistics New Zealand. *Linking Methodology Used by Statistics New Zealand in the Integrated Data Infrastructure Project* [Internet]. Wellington (NZ): Government of New Zealand; 2014 [cited 2017 Feb 18]. Available from: http://www.stats.govt.nz
14. Ministry of Health. *Ethnicity Data Protocols for the Health and Disability Sector* [Internet]. Wellington (NZ): Government of New Zealand; 2004 [cited 2017 Feb 18]. Available from: http://www.health.govt.nz
15. Atkinson J, Salmond C, Crampton P. *NZDep2013 Index of Deprivation*. Wellington (NZ): University of Otago Department of Public Health, Division of Health Sciences; 2014.
16. Statistics New Zealand. Unpublished Observations; 2017.
17. Ministry of Health. *Annual Update of Key Results 2015/16: New Zealand Health Survey* [Internet]. Wellington (NZ): Government of New Zealand; 2016 [cited 2017 May 6]. Available from: http://www.health.govt.nz
18. Jansen P, Bacal K, Crengle S. *He Ritenga Whakaaro: Māori Experiences of Health Services*. Auckland (NZ): Mauri Ora Associates; 2009.
19. Ministry of Health. *Primary Care Ethnicity Data Audit Toolkit: A Toolkit for Assessing Ethnicity Data Quality*. Wellington (NZ): Government of New Zealand; 2013.
20. Boyd M, Blakely T, Atkinson J. Ethnic counts on mortality, New Zealand Cancer Registry and Census Data: 2006-2011. *N Z Med J*. 2016;129(1429):22-39.
21. Gibb S, Das S. *Quality of Geographic Information in the Integrated Data Infrastructure* [Internet]. Wellington (NZ): Government of New Zealand; 2015 [cited 2017 Jan 18]. Available from: http://www.stats.govt.nz
22. Kukutai T, Taylor J. Data Sovereignty for Indigenous Peoples: Current Practice and Future Needs. In: *Indigenous Data Sovereignty: Toward an Agenda*. Research Monograph (Australian National University. Centre for Aboriginal Economic Policy Research); No.: 38. Canberra (AUST): ANU Press; 2016. p. 1-22.
23. Statistics New Zealand. *Experimental Population Estimates from Linked Administrative Data: Methods and Results* [Internet]. Wellington (NZ): Government of New Zealand; 2016 [cited 2017 Feb 8]. Available from: http://www.stats.govt.nz

Supporting Information

Additional supporting information may be found in the online version of this article:

Supplementary Table 1: The distribution of total and CVD populations by age and ethnicity for the IDI, HSU and 2013 Census populations.