Evidence for the use of an algorithm in resolving inconsistent and missing Indigenous status in administrative data collections

Daniel Christensen, Geoff Davis, Glenn Draper, Francis Mitrou, Sybille McKeown, David Lawrence, Daniel McAullay, Glenn Pearson, Wavne Rikkers, Stephen R. Zubrick

Abstract

Measures of the gap in living standards, life expectancy, education, health and employment between Indigenous and non-Indigenous Australians are primarily derived from administrative data sources. However, Indigenous identification in these data sources is affected by administrative practices, missing data, inconsistency, and error. As these factors have changed over time, assessing whether the gap between Indigenous and non-Indigenous Australians has changed over time, based on data unadjusted for these sources of error can potentially lead to misguided conclusions. Combining administrative data on the same individuals collected from different sources provides a method by which a more consistent derived Indigenous status can be applied across all records for an individual within a linked data environment. We used the Western Australian Data Linkage system to produce derived Indigenous statuses for individuals using a range of algorithms. We found that these algorithms reduced the amount of missing data and improved within-individual consistency. Based on these findings, we recommend our Multi-Stage Median algorithm be used as the standard indicator of Indigenous status for any reporting based on administrative datasets when multiple datasets are available for linkage, and that algorithmic approaches also be considered for improving the quality of other demographic variables from administrative data sources.

Keywords: data linkage, methodology, Indigenous inequality

Introduction

Administrative data collections maintained by Australian State, Territory, and Commonwealth human services agencies provide the basis for a range of key measures of social progress used in official reporting in Australia. These include measures used in the Council of Australian Governments (COAG) *Closing the Gap* reporting framework, designed to monitor the progress of Indigenous[1] Australians, relative to non-Indigenous Australians in living standards, life expectancy, education, health and employment. The Closing the Gap National Indigenous Reform Agreement has six major targets, including: closing the life expectancy gap within a generation; halving the gap in mortality rates for Indigenous children under five within a decade; and halving the gap for Indigenous students in reading, writing, and numeracy within a decade (COAG 2008). Measurement of progress against these targets depends upon the provision of administrative data.

Using administrative data for government reporting makes use of existing information that is a by-product of the day-to-day activities of human services agencies, rather than incurring the additional costs, resources and respondent burden that traditional survey methods impose (Holman et al. 2008). Despite the accepted benefits of using administrative data as a measurement tool, there are limitations, especially regarding consistent and reliable recording of the Indigenous status of individuals over time. Although Indigenous status in administrative collections in Australia is typically determined by a process of self-identification, Indigenous identification occurs in a complex cultural and historical context. For example, some individuals have reported that their willingness to self-identify as Indigenous is affected by perceptions of possible discrimination, and by the purpose of the data collection (ABS 2012b). Indigenous identification is also voluntary, so ‘missing’ or ‘not stated’ can be considered a valid response.

Historically, administrative data has been collected for the business needs of the respective custodian agencies, not for COAG reporting, and the quality and type of data collected has reflected this. Data collection practices can also vary between locations and across time, leading to further variation in data quality. In some settings the actual question regarding Indigenous identification may not be asked, and is either presumed by the collector or simply left unknown and missing. Another factor is the willingness of respondents, when asked, to provide personal information in a setting where they may feel uncomfortable about doing so. It remains the right of any individual to choose which Indigenous status they wish at the point of data collection, and that status may legitimately change from one identification occasion to the next. As a consequence of these issues, Indigenous status reporting on administrative data collections can be affected by missing, inconsistent and incorrect data. When multiple administrative data collections are statistically linked, these problems further increase the complexity of reporting, especially when an individual has differing Indigenous status recorded both within and across multiple data collections.

COAG Closing the Gap indicators are measures of the disparity between Indigenous and non-Indigenous Australians across a range of measures in living standards, life expectancy, education, health and employment. Without consistency in measurement of Indigenous status it becomes increasingly difficult to compare outcomes accurately across multiple time points, and misguided conclusions can be drawn from the data (Council 2012). Compositional changes in Indigenous and non-Indigenous populations can affect estimates of the magnitude of the gap. For example, although not an administrative data collection, results from the Australian Bureau of Statistics (ABS) 2011 Census of Population and Housing show substantial increases since the 2006 Census in the number of Indigenous persons in the Australian jurisdictions of Victoria (26.0 per cent), New South Wales (24.6 per cent) and the Australian Capital Territory (33.8 per cent), compared with a growth of 5.8 per cent in the Northern Territory (ABS 2012a). Furthermore, while the total Indigenous population grew 20.5 per cent between 2006 and 2011, the number of persons with an Indigenous status of 'not stated' still far exceeded the total Indigenous population, even though there was a 6.6 per cent decrease in the number of not-stated persons from 1,133,446 in 2006 to 1,058,586 in 2011. In as much as these population changes reflect either a changing propensity to identify as Indigenous and/or improvements in Census enumeration of Indigenous persons, assessing the disparity between Indigenous and non-Indigenous Australians without accounting for these compositional changes may lead to a false closing of the gap due to statistical artefacts rather than genuine progress.

Hunter and Ayyar (2011) compared several methods for accounting for the Indigenous undercount in NSW courts data. Their preferred Dual System Estimator approach effectively adjusts the population of all offenders upwards, based on a sample of offenders who have multiple independent ascertainments of Indigenous status. Their recommended approach increased Indigenous rates of offence from 119 to 243 offenders in every 1,000 Indigenous residents, again illustrating the impact of missing data on a measure of the gap between Indigenous and non-Indigenous Australians.

Despite the ramifications of these data quality limitations in Indigenous indicators, Australian governments and academic researchers are coming to rely more heavily on linked administrative data. Information flowing from administrative data is used to make important decisions around program spending and new policy initiatives that aim to deliver improvements in the lives of Indigenous Australians. Having the best quality data available to support decisions around program development and funding allocation is imperative.

One possible solution to missing, inconsistent and incorrect Indigenous identification within a linked data environment is to apply estimation algorithms that combine data according to specific rules (Draper et al. 2009; Taylor et al. 2012; Thompson, Woods & Katzenellenbogen 2012). For example, Taylor and colleagues (2012) examine the usage of algorithms to provide a means of correcting under-reporting of Indigenous mortality in the Australian state of New South Wales by linking death registration data for 2007 with four population health datasets relating to hospitalisations, emergency department attendances, and births. Reporting of deaths increased by 34.5 per cent using an algorithm based on a weight of evidence of a person being Indigenous, and by 56.6 per cent using an approach based on 'at least one report' of a person being Indigenous.

Our study seeks to demonstrate that the use of an appropriate algorithm can improve the consistency of Indigenous identification for individuals with records in linked administrative data collections beyond that which exists when only using their original records, and therefore support better reporting against State and Commonwealth progress indicators for Indigenous persons.

In contrast to the algorithmic solutions taken by others (Draper et al. 2009; Taylor et al. 2012; Thompson, Woods & Katzenellenbogen 2012), our study methodology:

- uses a greater variety of administrative collections from a range of human services agencies, over a longer period of time.
- uses a multiple linked administrative collection environment, where the records of any one individual are viewed across multiple data collections and over time to ascertain which algorithm gives the most consistent and reliable Indigenous status information.
- uses consistency as our key aim, which will be considered further in the Discussion section.

This paper draws on results from 'Getting Our Story Right – A cross agency data linkage and analysis project to better understand and improve information about Aboriginal and Torres Strait Islander peoples using administrative data collections'. This project was part-funded by COAG through the National Indigenous Reform Agreement (Schedule F – Data quality improvements) to provide evidence to inform 'National Best Practice Guidelines for Data Linkage Activities Relating to Aboriginal and Torres Strait Islander People' (AIHW & ABS 2012). This funding is in recognition of the issues caused by missing, inconsistent and incorrect Indigenous identification across time and across data collections. The algorithms in this paper may be considered an empirical test of the theory outlined in the AIHW–ABS Guidelines, which were developed as a result of the project (AIHW & ABS 2012).

Methods

Data collections

The data selected for this project consisted of eight administrative collections, with the collection and commencement year as follows: Birth Registrations, 1992; Communicable Disease, 1990; Emergency Data, 2002; Hospital Morbidity Data, 1970; Mental Health Information, 1967; Midwives Data, 1980; Death Registrations, 1969; and the Western Australian Literacy and Numeracy Assessment (WALNA), 1999.

We also used two survey collections: the Western Australian Aboriginal Child Health Survey (WAACHS), which covers 2000–2001 only; and the Health and Wellbeing Surveillance System, covering 2003 onwards. The WAACHS was a large-scale, state-wide, face-to-face cross-sectional survey of 1,999 households with 5,293 Indigenous children aged 0–17 years. This represented around one in six Indigenous children and young people in WA at the time (Silburn et al. 2006; Zubrick et al. 2004; 2005; 2006). The WA Health and Wellbeing Surveillance System (WAHWSS) is a telephone survey conducted on a sample of people of all ages who are currently resident in WA, with a monthly sample of at least 550 people (Tomlin, Joyce & Patterson 2012). Households are selected from the White Pages by a stratified random process.

Table 1 describes the ten collections considered in this project, which contributed to, and were adjusted by, the proposed algorithms.

Each of the ten collections included some version of Indigenous status. Whilst the column 'who collects Indigenous Identity data' describes the usual practice of obtaining data within each data collection, investigations as part of this project reveal considerable variation in actual practices. For example, anecdotal evidence, based on the authors' discussions with hospital administrators, suggests that in the often stressful setting of hospital admission procedures, the Indigenous status field may sometimes be populated from previous records for the patient in question, or a judgement may be made by the administrative officer completing the form based on physical appearance of the patient. Self-identification by individuals is in some cases impossible or impractical (for example, Death Registrations and WALNA respectively).

Birth Registration data currently record Indigenous status information for up to three people: the child, the mother, and the father. However, children's Indigenous status was not available to us at the time data was extracted for this project,[2] therefore the child's Indigenous status was derived from the parents' Indigenous status on this collection (decision process described in Table 2).

In contrast to the Birth Registration data, the Midwives data collection contains Indigenous status for the mother only – although the collection has since expanded to include the child's status field. We derived the child's Indigenous status from the mother's recorded Indigenous status. For the purposes of this project, because the father's Indigenous status is not collected on the form, the project team considered the father's Indigenous status to be missing. Following the logic set out in Table 2, where the mother is non-Indigenous the project team set the child's Indigenous status as 'missing' as there is deemed to be insufficient information by which to impute the child's Indigenous status for this project.

Ethical approval and Aboriginal representation

Ethical approval for this study was granted by the Department of Health Western Australia's Human Research Ethics Committee and the Western Australian Aboriginal Health Information and Ethics Committee.

Table 1: Data collections examined in this project

Collection	Information collected	Who collects Indigenous Identity data	Year commenced
Birth Registry Data	Births registered in WA.	Both parents are required to complete and sign the Birth Registration Form	1992
Communicable Diseases Notifiable and Infectious Disease Database (WANIDD)	Disease reports include all notifiable infectious diseases diagnosed in Western Australia, including both WA residents, and interstate or overseas visitors	Medical Practitioners, Nurse Practitioners, Pathologists	1990
Death Registrations Data	Deaths registered in WA (including cause and date)	Funeral Directors	1969
Health and Wellbeing Surveillance System	Telephone survey conducted on a monthly sample of at least 550 persons.	Edith Cowan Survey Research Centre	2003
Hospital Morbidity Data System (HMDS)	Separations from hospital and some details around nature of admission, illness, etc.	Hospital administrative staff	1970
Mental Health Information System (MHIS)	Patient separations from hospital and some details around nature of admission, illness, etc.	Hospital administrative staff	1967
Midwives Data Collection	Confinements registered in WA, only the mother's Indigenous status is reported	Midwives	1980
WA Emergency Database	Records date of presentation, presenting complaint, diagnosis, age, sex, Indigenous status, etc.	Clinicians in WA Emergency Departments	2002
Western Australian Aboriginal Child Health Survey (WAACHS)	A large-scale state-wide survey of 1,999 households with 5,293 Indigenous children aged 0–17 years	Household interviews by trained survey interviewers	2000–2001 only
Western Australian Literacy and Numeracy Assessment (WALNA)	Standardised testing of children at years 3,5 & 7 – this test is considered a precursor of the current NAPLAN testing	Teachers of the children being tested	1999

Table 2: Derivation of child's Indigenous status in Birth Registrations

Mother's Indigenous status	Father's Indigenous status	Child's derived Indigenous status
Indigenous	Indigenous	Indigenous
Indigenous	non-Indigenous	Indigenous
Indigenous	not stated/ unknown	Indigenous
non-Indigenous	Indigenous	Indigenous
non-Indigenous	non-Indigenous	non-Indigenous
non-Indigenous	not stated/ unknown	not stated/ unknown
not stated/ unknown	Indigenous	Indigenous
not stated/ unknown	non-Indigenous	not stated/ unknown
not stated/ unknown	not stated/ unknown	not stated/ unknown

The research team included Aboriginal representation to provide guidance on sensitive issues. The team also conducted stakeholder consultations with community members, data custodians, and Aboriginal researchers to gain further input into the project around self-identification and the development of a derived Indigenous status indicator.

Extraction, record linkage and merging of data

The data linkage was performed by the Western Australian Department of Health's Data Linkage Unit, which took no part in the analysis of linked data. The WA data linkage process involves probabilistic matching on demographic items and clerical review where required (Australia 2012; Holman et al. 2008). The process involved best practice separation of sensitive clinical and service data from demographic identifiers such that the data linkage repository itself contains a chain of linkages but no actual clinical and service data (Australia 2012). For more information on data linkage and extraction see the Western Australian Data Linkage Unit website (Australia 2012).

For the analysis, the team worked with de-identified data, with each individual represented by a project-specific encrypted identifier so that records for each individual from different collections could be compared. The master file contained all records for each individual and an Indigenous status flag ('Indigenous', 'non-Indigenous', 'missing') for each administrative contact that each individual had.

A total of 29,035,264 records were examined – after linkage, this was estimated to represent 4,189,958 individuals, with a median number of three contacts with the WA Data Linkage system. Of these records, 1,407,537 were affected by missing Indigenous status data and another 1,638,968 records were for people with inconsistent records – that is, were recorded as a mix of Indigenous and non-Indigenous across their records.

Algorithms to derive Indigenous status

A wide range of algorithms was initially considered in this project, including algorithms previously used by other researchers in this field (AIHW & ABS 2012; Draper et al. 2009; Taylor et al. 2012; Thompson, Woods & Katzenellenbogen 2012), as well as other algorithms unique to this project. After testing many different algorithms, the authors have focused on three algorithms which represent the breadth of possible results and best serve to demonstrate the utility of an algorithmic approach to missing, inconsistent, and incomplete Indigenous status data in administrative data collections.

The three algorithms were specified as follows:

- Ever Indigenous – If a person was coded as Indigenous at least once in any record, on any collection, they were assigned a derived status of 'Indigenous'. Of the algorithms investigated, this will result in the highest number of people coded as Indigenous.
- Always Indigenous – If a person was coded as Indigenous for every non-missing record on every collection, they were assigned

a derived status of 'Indigenous', otherwise they were coded as 'non-Indigenous' (if they had at least one record with an Indigenous status indicator of 'non-Indigenous') or 'missing' (if they only had records with missing Indigenous status fields). Of the algorithms investigated, this will result in the least number of people coded as Indigenous.

- The Multi-Stage Median algorithm – Each person was given a derived Indigenous status for each collection in which they have non-missing records, with the collection-derived Indigenous statuses combined into an overall derived Indigenous status for that person. Summing records at the collection level is a means of taking into account that some collections contain up to thousands of records for a person, while others contain only a single record. If this is not adjusted for, then an algorithm will only reflect status as recorded in whichever collection has the most records about a person. This may be problematic if we consider that records collected in different situations have greater independence than those collected within the same setting – there is some anecdotal evidence that records within collections tend to get 'rolled forward', which means that each extra record does not convey the same amount of extra information.

  The Multi-Stage Median algorithm is described in more detail in Appendix One, and is closely related to the AIHW–ABS (AIHW & ABS 2012) 'within and across data sets' method.

Other algorithms which were considered earlier in the analysis consisted of variations of a 50 per cent rule, in which a person's derived Indigenous status was defined as Indigenous if 50 per cent or more of their records were Indigenous. The Multi-Stage Median algorithm can be considered a more nuanced version of this algorithm. An attraction of the Multi-Stage Median algorithm over the 50 per cent rule is that in instances where a person has numerous records, but has only identified as Indigenous twice, the Multi-Stage Median algorithm still assigns a derived status of 'Indigenous'. Whilst this only leads to minor differences in aggregate statistics, this addresses the concern that individuals may choose to not identify as Indigenous under some circumstances due to fear of discrimination. This approach also reaches a compromise between the Ever and Always algorithms, which are sensitive to single erroneous records no matter how many records exist for a person.

These three algorithms were applied to 29,035,264 records across the ten collections. We compared the number of non-missing records for each person coded as either 'Indigenous' or 'non-Indigenous', and then calculated a derived Indigenous status of either 'Indigenous' or 'non-Indigenous', which was then applied to all records for that person, including those where the original Indigenous status indicator was 'missing'. The only records which resulted in a derived Indigenous status of 'missing' after the application of these algorithms were for those individuals where there was insufficient information for any of the algorithms to be applied – that is, where all their records had original Indigenous status indicators of ‘missing’. Following the application of an algorithm within the linked set of data extracts, each person was attributed an overall derived Indigenous status indicator. This new derived Indigenous status indicator was applied to all records for that person, and was calculated in addition to any original unadjusted fields for Indigenous status that already exist across each collection.

Analysis Methods

The analysis examined the selected collections to determine the impact of the three algorithms on the number of records classified as Indigenous, Non-Indigenous and Missing by collection.

Results

Comparison of original and algorithm-adjusted data collections

We compared original and algorithm-adjusted Indigenous statuses across all ten data collections, with the number of records in each category shown in Table 3.

Previously there were 1,638,968 records that were inconsistent at the person level – that is, the person had a mix of records classified as ‘Indigenous’, ‘non-Indigenous’ and ‘missing’, rather than all records belonging exclusively to one category. After application of the algorithms every person in the system had a consistent ‘derived’ indicator.

Each algorithm (Ever, Always and Multi-Stage Median) considered in this study had the same impact on the amount of missing data, such that the total number of records missing an Indigenous status flag across all collections was reduced from a total of 1,407,537 records out of 29,035,264 in the original data to 199,126 records after the application of each of the algorithms. This was a reduction of 86 per cent of missing Indigenous status. Those 199,126 records with missing data after the application of the algorithms reflected 170,337 people who had never been identified as either Indigenous or non-Indigenous – 152,457 of these people only had one record. This group had a median of one record per person compared to an overall median of three. This showed us that the majority of persons with an Indigenous status of ‘missing’ after the application of the algorithms were persons with relatively few records of contact with administrative collections.

Collection-specific decreases in missing statuses ranged from 34.8 per cent (Hospital Morbidity) to 100 per cent (WAACHS, WALNA). The midwives’ records for mothers were unadjusted as there were no missing data in the original series. The greatest absolute decrease in missing data was for midwives’ records for children, a decrease of 706,749 records (95.7 per cent). However, this decrease pertains to the way in which we derived children’s status from the original midwives collection (see ‘Data Collections’ section), rather than to a problem with the collection.

Although each algorithm reduced the amount of missing statuses by the same amount, the impact of the algorithms on Indigenous counts varied by algorithm and collection. For the Ever algorithm, the greatest proportional increase in Indigenous status occurred in the deaths registrations (68 per cent greater than original) and in the WAHWSS (62.6 per cent). For the Multi-Stage Median algorithm, the greatest proportional increase in counts of Indigenous status occurred in the deaths registrations (34.8 per cent greater than original) and in midwives data for babies (21.5 per cent). The Multi-Stage Median algorithm led to slight decreases in the count of Indigenous people in the WAACHS, WAHWSS and WALNA.

The impact of the three algorithms examined varied by the original data collection they were applied to and the metric used to measure change. Death registrations from 1970–2010 showed the largest proportional increase in Indigenous identifications from the Multi-Stage Median and Ever algorithms (34.8 per cent and 68.0 per cent respectively); this increase partially reflects the algorithms' ability to assign Indigenous status prior to the official collection of Indigenous status in this collection in 1983. The largest numerical changes in Indigenous identifications following the application of the Multi-Stage Median and Ever algorithms were in the Hospital Morbidity and Emergencies datasets.

Table 3: Records in the Western Australian Data Linkage System at the time of extraction*

Collection	Indigenous	Non-Indigenous	Missing	Total
Births registrations – baby, original series	34,023 (6.7%)	427,602 (84.0%)	47,126 (9.3%)	508,751
Births registrations – baby, Ever algorithm	40,225 (7.9%)	462,770 (91.0%)	5,756 (1.1%)	508,751
Births registrations – baby, Always algorithm	22,873 (4.5%)	480,122 (94.4%)	5,756 (1.1%)	508,751
Births registrations – baby, Multi-Stage Median algorithm	34,778 (6.8%)	468,217 (92%)	5,756 (1.1%)	508,751
Births registrations – father, original series	22,647 (4.5%)	434,770 (86.8%)	43,671 (8.7%)	501,088
Births registrations – father, Ever algorithm	28,570 (5.7%)	453,606 (90.5%)	18,912 (3.8%)	501,088
Births registrations – father, Always algorithm	11,491 (2.3%)	470,685 (93.9%)	18,912 (3.8%)	501,088
Births registrations – father, Multi-Stage Median algorithm	23,280 (4.6%)	458,896 (91.6%)	18,912 (3.8%)	501,088
Births registrations – mother, original series	26,663 (5.3%)	445,506 (88.9%)	28,919 (5.8%)	501,088
Births registrations – mother, Ever algorithm	38,378 (7.7%)	462,467 (92.3%)	243 (0%)	501,088
Births registrations – mother, Always algorithm	11,482 (2.3%)	489,363 (97.7%)	243 (0%)	501,088

Births registrations – mother, Multi-Stage Median algorithm	28,477 (5.7%)	472,368 (94.3%)	243 (0%)	501,088
Communicable diseases, original series	47,066 (18.5%)	140,797 (55.3%)	66,737 (26.2%)	254,600
Communicable diseases, Ever algorithm	58,058 (22.8%)	186,016 (73.1%)	10,526 (4.1%)	254,600
Communicable diseases, Always algorithm	37,613 (14.8%)	206,461 (81.1%)	10,526 (4.1%)	254,600
Communicable diseases, Multi-Stage Median algorithm	54,194 (21.3%)	189,880 (74.6%)	10,526 (4.1%)	254,600
Death registrations, original series	10,424 (2.5%)	281,595 (68.7%)	117,894 (28.8%)	409,913
Death registrations, Ever algorithm	17,514 (4.3%)	344,903 (84.1%)	47,496 (11.6%)	409,913
Death registrations, Always algorithm	8,482 (2.1%)	353,935 (86.3%)	47,496 (11.6%)	409,913
Death registrations, Multi-Stage Median algorithm	14,048 (3.4%)	348,369 (85.0%)	47,496 (11.6%)	409,913
Health and Wellbeing Surveillance System, original series	796 (2.1%)	34,552 (92.4%)	2,036 (5.4%)	37,384
Health and Wellbeing Surveillance System, Ever algorithm	1,294 (3.5%)	35,990 (96.3%)	100 (0.3%)	37,384
Health and Wellbeing Surveillance System, Always algorithm	258 (0.7%)	37,026 (99%)	100 (0.3%)	37,384
Health and Wellbeing Surveillance System, Multi-Stage Median algorithm	717 (1.9%)	36,567 (97.8%)	100 (0.3%)	37,384
Hospital Morbidity Data System, original series	1,186,483 (6.6%)	16,801,195 (92.9%)	102,269 (0.6%)	18,089,947
Hospital Morbidity Data System, Ever algorithm	1,580,386 (8.7%)	16,442,890 (90.9%)	66,671 (0.4%)	18,089,947
Hospital Morbidity Data System, Always algorithm	643,355 (3.6%)	17,379,921 (96.1%)	66,671 (0.4%)	18,089,947
Hospital Morbidity Data System, Multi-Stage Median algorithm	1,299,755 (7.2%)	16,723,521 (92.4%)	66,671 (0.4%)	18,089,947
Mental Health Information System, original series	11,623 (4.4%)	240,105 (90.6%)	13,365 (5%)	265,093
Mental Health Information System, Ever algorithm	17,978 (6.8%)	245,377 (92.6%)	1,738 (0.7%)	265,093
Mental Health Information System, Always algorithm	5,442 (2.1%)	257,913 (97.3%)	1,738 (0.7%)	265,093
Mental Health Information System, Multi-Stage Median algorithm	12,988 (4.9%)	250,367 (94.4%)	1,738 (0.7%)	265,093
Midwives data collection – babies, original series	45,672 (5.8%)	–	738,447 (94.2%)	784,119
Midwives data collection – babies, Ever algorithm	64,402 (8.2%)	688,019 (87.7%)	31,698 (4.0%)	784,119

Midwives data collection – babies, Always algorithm	36,091 (4.6%)	716,330 (91.4%)	31,698 (4.0%)	784,119
Midwives data collection – babies, Multi-Stage Median algorithm	55,490 (7.1%)	696,931 (88.9%)	31,698 (4.0%)	784,119
Midwives data collection – mother, original series	45,195 (5.8%)	727,963 (94.2%)	–	773,158
Midwives data collection – mother, Ever algorithm	62,393 (8.1%)	710,765 (91.9%)	–	773,158
Midwives data collection – mother, Always algorithm	20,183 (2.6%)	752,975 (97.4%)	–	773,158
Midwives data collection – mother, Multi-Stage Median algorithm	47,806 (6.2%)	725,352 (93.8%)	–	773,158
WA emergencies database, original series	799,114 (12.5%)	5,343,706 (83.6%)	247,048 (3.9%)	6,389,868
WA emergencies database, Ever algorithm	985,954 (15.4%)	5,387,928 (84.3%)	15,986 (0.3%)	6,389,868
WA emergencies database, Always algorithm	485,696 (7.6%)	5,888,186 (92.1%)	15,986 (0.3%)	6,389,868
WA emergencies database, Multi-Stage Median algorithm	871,858 (13.6%)	5,502,024 (86.1%)	15,986 (0.3%)	6,389,868
Western Australian Aboriginal Child Health Survey, original series	7,212 (93.6%)	475 (6.2%)	19 (0.2%)	7,706
Western Australian Aboriginal Child Health Survey, Ever algorithm	7,290 (94.6%)	416 (5.4%)	–	7,706
Western Australian Aboriginal Child Health Survey, Always algorithm	3,856 (50.0%)	3,850 (50.0%)	–	7,706
Western Australian Aboriginal Child Health Survey, Multi-Stage Median algorithm	7,033 (91.3%)	673 (8.7%)	–	7,706
Western Australian Literacy and Numeracy Assessment, original series	30,206 (5.9%)	482,337 (94.1%)	6 (0.0%)	512,549
Western Australian Literacy and Numeracy Assessment, Ever algorithm	39,327 (7.7%)	473,222 (92.3%)	–	512,549
Western Australian Literacy and Numeracy Assessment, Always algorithm	15,979 (3.1%)	496,570 (96.9%)	–	512,549
Western Australian Literacy and Numeracy Assessment, Multi-Stage Median algorithm	29,013 (5.7%)	483,536 (94.3%)	–	512,549
Total, original series	2,267,124 (7.8%)	25,360,603 (87.3%)	1,407,537 (4.8%)	29,035,264
Total, Ever algorithm	2,941,769 (10.1%)	25,894,369 (89.2%)	199,126 (0.7%)	29,035,264
Total, Always algorithm	1,302,801 (4.5%)	27,533,337 (94.8%)	199,126 (0.7%)	29,035,264
Total, Multi-Stage Median algorithm	2,479,437 (8.5%)	26,356,701 (90.8%)	199,126 (0.7%)	29,035,264

* total number of records

In every instance, the Always algorithm decreased Indigenous counts compared with original data. The greatest proportional decrease in Indigenous status occurred in the WAHWSS (67.6 per cent) and birth registrations data for mothers (56.9 per cent).

Discussion

This study examined the impact of using an algorithmic approach to improve missing and inconsistent Indigenous status data in an administrative linked data environment. We found that applying algorithms to derive Indigenous status made a substantial difference to a range of health and development indicators. Based on these findings, we recommend our Multi-Stage Median algorithm be used as the standard indicator of Indigenous status for any reporting based on administrative datasets when multiple datasets are available for linkage, and that algorithmic approaches also be considered for improving the quality of other demographic variables from administrative data sources.

The three algorithms examined in this paper substantially reduced the number of records missing an Indigenous status indicator, such that the number of records missing an Indigenous status flag was reduced from a total of 1,407,537 records out of 29,035,264 in the original data to 199,126 records after the application of the algorithms. This 86 per cent reduction in missing Indigenous status across all records considered is important, as it has direct implications for government policy and decisions based on administrative data. Given the investment of public resources by government, it is appropriate that decisions are based on a methodology which is demonstrably better than the methodologies currently used.

The three algorithms resulted in consistency across time and collections for any given individual – that is, using an algorithm resulted in a person being always counted the same way in each year and in each administrative context. This may ameliorate the impact of recent trends showing an increasing propensity to identify as Indigenous.

Whilst all of the algorithms reduced the amount of missing data and increased internal consistency for individuals across all their records, the recommended Multi-Stage Median algorithm has additional benefits that the other algorithms do not possess. This algorithm produced results robust against error in any individual collection by combining all available Indigenous status information about a given individual within the data linkage system and finding a plausible derived status, whilst the Ever and Always algorithms are sensitive to single misidentifications in either direction.

The Multi-Stage Median algorithm produced results broadly congruent with the range of adjustments suggested by published under-identification studies (AIHW 2013; Lawrence et al. 2012). The AIHW has estimated Indigenous under-identification in hospital separations data using a follow-up survey methodology. The recent AIHW report (2013) recommends a correction factor of 1.01 (one per cent) across hospital separations in Western Australia and a 2010 report recommended a three per cent correction factor. Comparisons by

Lawrence and colleagues (2012) between the WAACHS and midwives data suggest that the Indigenous count in the midwives notification system would have been approximately 11 per cent higher had it been collected at the same time and in the same manner as the WAACHS.

Other methods have also been suggested to address quality concerns with Indigenous identification data (ABS 2008; AIHW 2012; Taylor et al. 2012). The ABS linked death registrations between 9 August 2006 to 30 June 2007 to Census data (8 August 2006) as part of the Census Data Enhancement Indigenous Mortality Quality Study (ABS 2008). Although there are differences between the Census and administrative data sets, the probabilistic approach taken to linkage by the ABS is similar to methods used in other linkage studies referenced in this paper, and the Census gives another useful source by which to adjust administrative data. However, while this method gives an estimate of under-enumeration in Indigenous deaths by jurisdiction, the Census Data Enhancement does not lend itself to extension beyond the specific circumstances under which it was developed. For example, it is unclear how these adjustments would translate into: an analysis of cause of death in inter-censal years; other administrative sources such as Births or Education; and/ or a linked data analysis looking at the association between hospitalisation and subsequent mortality. The AIHW undertook a similar analysis of deaths between 2001 and 2006 (AIHW 2012), linking death registrations to the National Hospital Morbidity Data set, the Residential Aged Care data set, and the National Perinatal Data Collection (NPDC). The AIHW analysis reported similar adjustments to the ABS Census Data Enhancement project. While this approach is potentially sustainable in inter-censal periods, it is unclear how it would extend more broadly to other collections or topics.

The approach used in this paper has similarities to the approach taken by Taylor and colleagues (2012) in their analysis on adjusting death registration data, but with a crucial difference. Taylor and colleagues'algorithm includes a 'default to original if Indigenous' rule. That is, if a person's death record has them coded as Indigenous, their death is given a derived status of 'Indigenous', regardless of any other administrative contacts the person may have had. If this approach is extended beyond adjusting one collection at a time the 'default to original' rule leads to different Indigenous statuses in different collections. In contrast, the multi-stage median algorithm results in a methodology which can be applied consistently across all records for an individual within any data linkage system.

Limitations

An under-explored potential limitation is the sensitivity of algorithms to different data environments, such as the number of data collections used, the duration over which these collections span, and other issues such as error rates. We have undertaken a preliminary analysis of these issues in Appendix Two, but recognise that more work needs to be undertaken between jurisdictions to resolve this issue fully. Areas of specific interest include the impact of changing administrative practices over time and by geography (especially remoteness), and whether the number of records an individual has leads to any bias in the approach. For example, where an individual has records spanning a time frame including the 1960s, 1970s and 1980s, do changes to the method of Indigenous status collection over that period bias the measures derived through our method? Similarly, although the project team recommends that our approach could be used in other data-linkage environments for variables other than Indigenous status, this option remains unexplored.

We have not explored in this paper whether the proposed algorithm systematically identifies individuals in a way that is likely to affect any rates or analyses subsequently derived; that is, whether there is some sort of selection bias affecting which individuals' records are adjusted, and if so, whether this pattern of adjustment seems plausible. For example, does the algorithm disproportionately affect individuals with multiple administrative contacts, and if so, is this adjustment appropriate? We acknowledge this as an area for future work. However, it is our contention that a method of adjustment which is based on each individual's own pattern of identification, internally consistent, and has a plausible overall pattern of adjustment provides a more sound basis for estimation than the present situation whereby analysts are forced to work with original – unadjusted – data or ad hoc adjustment methods.-

The guiding principle adopted by the project team is that as more data collections are made available they should be added to our algorithmic approach: a process of continuous improvement. A consequence of this approach is that as new records are added for an individual their derived Indigenous status may change. That is, Indigenous estimates developed from algorithms can potentially change as additional data becomes available. Although this feature appears initially unattractive, it is a necessary consequence of any approach that assigns a consistent Indigenous identifier over time, and it is our contention that it is difficult to make a case against continuous improvement in data quality.

The approach in this paper deliberately steps away from the idea of 'most trusted' or 'gold standard' measures of Indigenous status. There are many factors that can affect the recording of an individual's Indigenous status, one of them being individual choice. The principle of self-identification explicitly includes the right of individuals to choose to identify differently under different circumstances. The view we developed in stakeholder consultation and in discussion within the project team is that if a person chooses to identify as Indigenous twice and non-Indigenous once, we would not say that any of their identifications are 'wrong' or 'less trusted'. Accordingly, we focussed on an approach that was internally consistent, avoiding discussions of more or less-trusted datasets or gold standards, which we felt were inconsistent with the principle of self-identification. We acknowledge that the views of the project team are not universally accepted in this sphere, and that further consultation may be required. We also acknowledge that continuing work investigating and improving the quality of administrative collections needs to ensue.

The approach presented in this paper is explicitly aimed at a linked data environment, and does not represent a viable solution in single-collection situations for persons who only have one record.

Finally, our proposed method cannot be considered in isolation. Indigenous identity and identification reflects a complex interplay of individual and contextual factors. Any data user working with recorded Indigenous status needs to be aware of this context.

Conclusions

A benefit of our algorithmic approach is that it produces a consistent derived identifier when conducting analyses using more than one record per person – for instance, either a series of hospital records for a person, or a study looking at the impact of birth conditions on subsequent educational outcomes for a child. Previously, individual analysts would be left to make their own decisions on how to combine the Indigenous status across records, effectively an ad hoc algorithmic approach that only made use of the records within the datasets of interest, rather than relying on the combined power of an entire Data Linkage System. With the rapid growth in the use of linked data in research and government reporting, our proposed methodology has the potential to increase the quality of reporting that is required to interpret and apply the data of administrative collections.

The algorithmic approach is important because it has positive implications for both policy and reporting and analysis. In the Australian context, the COAG *Closing the Gap* initiative aims to reduce inequalities between Indigenous and non-Indigenous Australians in living standards, life expectancy, education, health and employment. Reflecting a considerable investment of resources, these policy decisions need to be made on a consistent, transparent, and accurate basis. At present COAG indicators rely on unadjusted Indigenous status as recorded in administrative data ('original' data), which is affected by missing and inconsistent identification. As this study shows, unadjusted Indigenous status may lead to an under-enumeration of Indigenous persons and an under-reporting of key outcomes such as deaths and communicable diseases. Whilst our approach does not completely eliminate the impact of a changing propensity to identify, by assigning a consistent identifier to each individual, we increase comparability across time. The algorithmic approach is also dynamic, in that the adjustments produce change as new data come to hand, resulting in continuous quality improvement of the derived indicator with each successive application of the algorithm.

This study demonstrates that after taking into account algorithmic corrections to derive consistent identification between data collections we see more deaths, more communicable diseases and more hospital separations for Indigenous people in Western Australia. Such an adjustment has serious policy and practice implications, and calls into question the continued use of original unadjusted data in the reporting of *Closing the Gap* indicators. We recommend our Multi-Stage Median algorithm be used as the standard indicator of Indigenous status for any reporting based on administrative datasets when multiple datasets are available for linkage. This is particularly important as we move to an environment of multiple linked data collections for more sophisticated COAG reporting.

As an extension of this project the WA Data Linkage Unit is planning to prepare a regularly updated derived Indigenous status indicator using the Multi-Stage Median algorithm as presented in this paper. This indicator – along with instructions for use – will be available on request to future research projects. Although our study was conducted in Western Australia, data linkage is growing both within individual jurisdictions and at a national level, with every other state also having a linkage program in place. The proposed Multi-Stage Median algorithm has the potential to improve Indigenous statistics across all jurisdictions. More broadly, this paper reveals a method for combining data within an administrative data context wherever multiple ascertainments of a variable exist, and there remain issues with missing, inconsistent and incorrect data.

Competing Interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed to the development of the study methodology. Glenn Draper acquired and analysed the data, Daniel Christensen and Geoff Davis wrote the first draft of the manuscript. All authors edited the paper. All authors read and approved the final manuscript.

Acknowledgements

The views expressed in this paper are those of the authors and not those of any of the funding agencies. These agencies are not responsible for the analyses, interpretations or conclusions reached by the authors.

We thank the Western Australian Department of the Attorney General, the Western Australian Department of Education, the Telethon Kids Institute, and the Western Australian Department of Health for providing the data which allowed this project to be undertaken. We also thank Dr. Lee Taylor from the New South Wales Ministry of Health for her advice around the development of the algorithms in this project.

This project was part-funded by the ABS and part-funded by COAG through the National Indigenous Reform Agreement (Schedule F – Data quality improvements) to provide evidence supporting 'National Best Practice Guidelines for Data Linkage Activities Relating to Aboriginal and Torres Strait Islander People' (AIHW & ABS 2012).

Endnotes

[1] 'Indigenous Australians' refers to persons of Aboriginal and/or Torres Strait Islander descent.

[2] It is important to note that information on the child's Indigenous status is now collected and will be available for subsequent research.

References

ABS (Australian Bureau of Statistics) (2008) Discussion Paper: Assessment of Methods for Developing Life Tables for Aboriginal and Torres Strait Islander Australians, 2006, Cat. No. 3302.0.55.002, ABS Canberra.

—— (2012a) Census of Population and Housing – Counts of Aboriginal and Torres Strait Islander Australians, 2011, Cat. No. 2075.0, ABS, Canberra.

—— (2012b) Information paper – Perspectives on Aboriginal and Torres Strait Islander Identifications in Selected Data Collection Contexts, Australia 2012, Cat. No. 4726.0, ABS, Canberra.

AIHW (Australian Institute of Health and Welfare) (2012) An enhanced mortality database for estimating Indigenous life expectancy. A feasibility study, 2012, IHW 75, Canberra, AIHW.

—— (2013) Indigenous identification in hospital separations data – quality report, IHW 90, AIHW, Canberra.

AIHW & ABS (2012) National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people, vol. IHW 74, AIHW, Canberra.

COAG (Council of Australian Governments) (2008) National Indigenous Reform Agreement (Closing the Gap), COAG, Canberra.

—— (2012) Indigenous reform 2010–11: Comparing performance across Australia, COAG Reform Council, Sydney.

Data Linkage Western Australia (2012) Linkage Process, www.datalinkage-wa.org.au (accessed 24 September 2014).

Draper, G.K., Somerford, P.J., Pilkington, A.A.G. & Thompson, S.C. (2009) 'What is the impact of missing Indigenous status on mortality estimates? An assessment using record linkage in Western Australia', *Australian and New Zealand Journal of Public Health*, 33 (4), 325–331.

Holman, C.A.J., Bass, A., Rosman, D.L., Smith, M.B., Semmens, J.B., Glasson, E.J., Brook, E.L., Trutwein, B., Rouse, I.L., Watson, C.R., de Klerk, N.H. & Stanley, F.J. (2008) 'A Decade of Data Linkage in Western Australia: Strategic Design, Applications and Benefits of the WA Data Linkage System', *Australian Health Review*, 32 (4), 766–777.

Hunter, B. & Ayyar, A. (2011) 'Undercounts in Offender Data and Closing the Gap Between Indigenous and Other Australians', *The Australian Journal of Social Issues*, 46 (1), 69–90.

Lawrence, D., Christensen, D., Mitrou, F., Draper, G., Davis, G., McKeown, S., McAullay, D., Pearson, G. & Zubrick, S.R. (2012) 'Adjusting for under-identification of Aboriginal and/or Torres Strait Islander births in time series produced from birth records: Using record linkage of survey data and administrative data sources', *BMC Medical Research Methodology*, 12 (90).

Silburn, S.R., Zubrick, S.R., De Maio, J.A., Shepherd, C., Griffin, J.A., Mitrou, F.G., Dalby, R.B., Hayward, C. & Pearson, G. (2006) The Western Australian Aboriginal Child Health Survey: strengthening the capacity of Aboriginal children, families and communities, Perth, Curtin University of Technology and Telethon Institute for Child Health Research.

Taylor, L.K., Bentley, J., Hunt, J., Madden, R., McKeown, S., Brandt, P. & Baker, D. (2012) 'Enhanced reporting of deaths among Aboriginal and Torres Strait Islander peoples using linked administrative health datasets', *BMC Medical Research Methodology*, 12 (91).

Thompson, S.C., Woods, J.A. & Katzenellenbogen, J.M. (2012) 'The quality of Indigenous identification in administrative health data in Australia: insights from studies using data linkage', *BMC Medical Informatics and Decision Making*, 12 (91).

Tomlin, S., Joyce, S. & Patterson, C. (2012) Health and Wellbeing of Adults in Western Australia 2011, Overview and Trends, Department of Health, Western Australia.

Zubrick, S.R., Lawrence, D.M., Silburn, S.R., Blair, E., Milroy, H, Wilkes, T, Eades, S, D'Antoine, H, Read, A, Ishiguchi, P & Doyle, S. (2004) The Western Australian Aboriginal Child Health Survey: The Health of Aboriginal Children and Young People, Perth, Telethon Institute for Child Health Research.

Zubrick, S.R., Silburn, S.R., De Maio, J.A., Shepherd, C., Griffin, J., Dalby, R.B., Mitrou, F.G., Lawrence, D.M., Hayward, C., Pearson, G., Milroy, H., Milroy, J. & Cox, A. (2006) The Western Australian Aboriginal Child Health Survey: improving the educational experiences of Aboriginal children and young people, Perth, Curtin University of Technology and Telethon Institute for Child Health Research.

Zubrick, S.R., Silburn, S.R., Lawrence, D.M., Mitrou, F.G., Dalby, R.B., Blair, E.M., Griffin, J., Milroy, H., De Maio, J.A., Cox, A. & Li, J. (2005) The Western Australian Aboriginal Child Health Survey: the social and emotional wellbeing of Aboriginal children and young people, Perth, Curtin University of Technology and Telethon Institute for Child Health Research.

Appendix One: The Multi-stage Median Algorithm methodology

In the Multi-Stage Median algorithm each person is given a derived Indigenous status for each collection they have non-missing records in, with the collection-derived Indigenous statuses combined into an overall derived Indigenous status for that person.

The algorithm follows a number of steps, first separately for every collection (steps a–d), then combined across all collections (steps e–h):

- Missing records within a collection cannot be used to determine a derived Indigenous status. A person who only has records with missing Indigenous status within a collection is given a collection-derived status of missing for that collection.
- If a person has only one non-missing record within a collection, this becomes their derived Indigenous status for that collection.
- If a person has a total of two non-missing records within a collection, and at least one identifies them as Indigenous, then that person is given a derived status of Indigenous for that collection. Otherwise, if both records have a status of non-Indigenous, then that person is assigned a derived status of non-Indigenous.
- If a person has a total of three or more non-missing records within a collection, and two or more of these are Indigenous, then that person is given a derived status of Indigenous for that collection. Otherwise, if the person has three or more non-missing records and two or more are recorded as non-Indigenous then that person is assigned a derived status indicator of non-Indigenous.
  This gives a person a derived Indigenous status for each collection in which they have records. This process is then repeated with the derived status for each collection:
- Collections with collection-derived statuses of missing cannot be used to determine an overall derived Indigenous status. If a person only has records in collections with collection-derived statuses of missing, then that person is given an overall derived status of missing.
- If a person has only one non-missing collection-derived status, this becomes their overall derived Indigenous status.
- If records for a person occur in a total of two non-missing collections and at least one has a collection-derived status of Indigenous, then that person is given an overall derived status of Indigenous. Otherwise, if both collections have collection-derived non-Indigenous status flags, then the overall derived status is set as non-Indigenous.
- If records for a person occur in a total of three or more non-missing collections, and two or more of these have a collection-derived status of Indigenous, then that person is given an overall derived status of Indigenous. Otherwise, the person is assigned an overall derived status indicator of non-Indigenous.

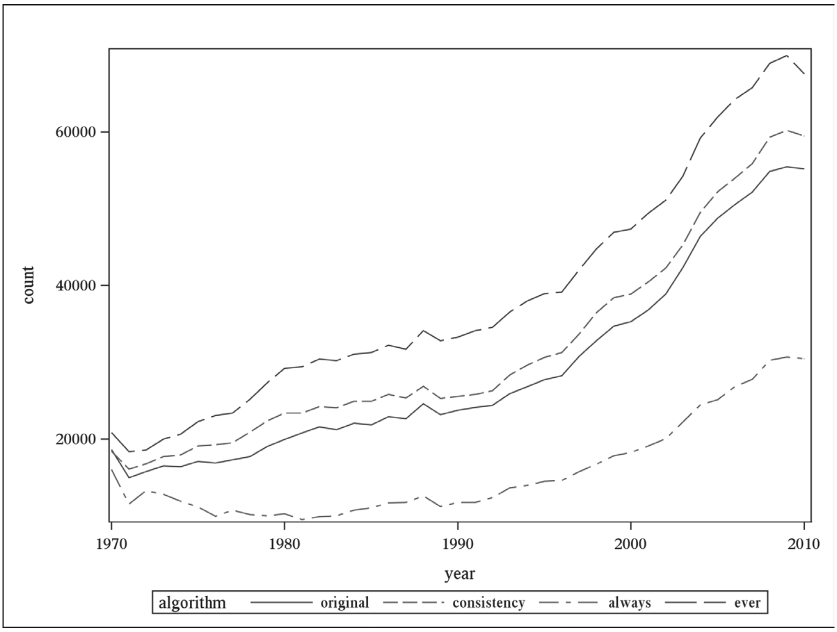
Appendix Two: The impact of additional collections

To examine the impact of additional collections on the Multi-Stage Median algorithm, the number of Indigenous separations on the Hospital Morbidity Data System was estimated under the following conditions:

- all ten data collections;
- all data collections except for morbidity (nine collections);
- morbidity, births, deaths and midwives collections (four collections); and
- morbidity alone (original data; one collection).

Figure 1: Indigenous hospital separations, WA, 1970–2010 by various input collections

The ten-collection and four-collection calculations produced similar change compared with the original series – from -1.3 per cent in 1970 to 18.1 per cent in 1978 with an average change of 10.1% for the ten-collection algorithm and -1.1% in 1970 to 19.2% in 1978 with an average change of 12.2% for the 4-collection algorithm).



The nine-collection algorithm, which estimates hospital separations without using any Indigenous identifications collected within the hospital separations system represents an *a fortiori* demonstration of the algorithmic approach, as it discards each person's original Indigenous identification within the hospital morbidity data system.