Examining the quality of name code record linkage: what is the impact on death and cancer risk estimates? A validation study

Alexander Swart,[1] Nicola S. Meagher,[1] Marina T. van Leeuwen,[1] Kun Zhao,[2] Andrew Grulich,[3] Limin Mao,[4] Deborah Anne Randall,[5] Louisa Degenhardt,[6,7] Lucy Burns,[6] Dianne O'Connell,[8] Janaki Amin,[3] Claire M. Vajdic[1]

Abstract

Objective: To examine the validity and impact of record linkage using name code compared to full name records.

Methods: A registry of 45,419 opioid substitution clients (1985–2007) was linked with national population-based death and cancer registries using registrant's name, date of birth, sex, state, postcode and date of death. Records were linked using full name and then using the first two letters of the given and surname (2x2 name code). Sensitivity and specificity were computed and regression analysis used to identify factors related to linkage accuracy. Standardised mortality ratios (SMR) and standardised cancer incidence ratios (SIR) were estimated.

Results: The sensitivity and specificity of name code compared to full name linkage were 65.31% and 99.91% for death records and 76.81% and 99.89% for cancer records. Registrants' age and sex and accuracy of the registries were associated with risk of false linkages. Death and cancer risks (SMR 6.98, 95%CI 6.77–7.19; SIR 1.16, 95%CI 1.08–1.24) were significantly underestimated using name code linkage (SMR 4.39, 95%CI 4.23–4.56; SIR 0.92, 95%CI 0.85–0.99).

Conclusion: Record linkage using 2x2 name code has low sensitivity but high specificity, resulting in conservative estimates of death and cancer risk. This may translate to meaningful differences in outcomes.

Key words: record linkage, name code, accuracy, validity, sensitivity, specificity, mortality, cancer, risk

Record linkage is an established method of discovering and quantifying disease trends, associations and health service utilisation.[1-3] In particular, record linkage between disease registries and administrative population-based health databases is an efficient and powerful research tool to investigate the medium- and long-term risks of cancer and death in a population.[4-7]

Most countries do not have a unique, lifetime person-level identifier available to facilitate deterministic, anonymous record linkage between databases. Probabilistic linkage algorithms based on personal identifiers such as name, sex, and date of birth are therefore necessary, and give the relative likelihood that two records on different databases relate to the same individual. Probabilistic record linkage between good quality databases can achieve high linkage accuracy, for example false-positive and false-negative match rates of 1/1000 records.[8] Due to privacy concerns and in some instances legislative constraints, some disease registries (particularly those for which there may be concerns over stigma related to being identified with such a condition, such as HIV or sexually transmitted infections) record only an individual's 'name code', for example, the first two letters of the registrant's first name and surname.[2,9,10] However, this method may compromise the utility of record linkage with these registries.

Linkage using records containing truncated, absent, or ambiguous name information has been shown to have high sensitivity (82% to 96%) and specificity (92% to 100%).[11-16] These studies investigated factors contributing to linkage accuracy by modifying the linkage algorithm, and generally concluded that probabilistic matching of such records was appropriate and effective, and that resultant measures of disease incidence or association were reliable. However, none of these studies directly compared name code linkage to full name linkage.

We conducted a validation study of name code (first two letters of first and surname) compared to full name record linkage between population-based administrative databases. Linkage sensitivity and specificity

1. Adult Cancer Program, Prince of Wales Clinical School, University of New South Wales
2. Cancer and Screening Unit, Australian Institute of Health and Welfare, ACT
3. The Kirby Institute, University of New South Wales
4. National Centre in HIV Social Research, University of New South Wales
5. Centre for Health Research, University of Western Sydney, New South Wales
6. National Drug and Alcohol Research Centre, University of New South Wales
7. Melbourne School of Population and Global Health, University of Melbourne, Victoria
8. Cancer Research Division, Cancer Council NSW
Correspondence to: Associate Professor Claire M. Vajdic, Adult Cancer Program, Lowy Cancer Research Centre, University of New South Wales, NSW 2052; e-mail: claire.vajdic@unsw.edu.au
Submitted: April 2014; Revision requested: July 2014; Accepted: July 2014
The authors have stated they have no conflict of interest.

Aust NZ J Public Health. 2015; 39:141-7; doi: 10.1111/1753-6405.12287

were estimated, the relationship between cohort characteristics and linkage accuracy examined, and the mortality and cancer risk estimates obtained via linkage using name code and full name compared.

Methods

Study population and databases

The study population included all adults (≥16 years) registered for opioid substitution therapy (OST) in New South Wales (NSW), Australia, in 1985–2007 (n=45,482). The Pharmaceutical Drugs of Addiction System (PHDAS) is a fully-identified administrative database of individuals receiving methadone or buprenorphine under the OST program in NSW. Prior to starting therapy, an individual must show proof of identity including name and date of birth to the authorised prescribing doctor.[17] Some registrants (14%) had multiple names recorded in the database, such as an alias, a maiden or married surname, or the use of a middle name as a first name. Registrants were excluded from analysis where essential variables for record linkage were missing (e.g. date of birth; n=12, 0.03%), or where dates were inconsistent (e.g. OST end date preceded OST start date; n=51, 0.11%), resulting in a cohort of 45,419 registrants. The median age of the OST registrants was 27 years at first entry into OST (interquartile range 23–32).

The National Death Index (NDI) is a database containing full names of all individuals who have died in Australia from 1980. Prior to 1997, the database contained only the year of birth for deaths registered in some of the jurisdictions (referred to herein as 'early NDI') but from 1997 the full date of birth was recorded nationally (referred to herein as 'later NDI'). The Australian Cancer Database (ACD) is a fully identified database of all invasive primary cancers (except squamous and basal cell carcinoma of the skin) notified by mandate to the jurisdictional cancer registries since 1982.

Record linkage: full name

The following variables were used for the full name record linkage: first name, middle name, and surname for the primary name and any other recorded name, date of birth, gender, date of last contact (i.e. the last OST program start date on the PHDAS), state of residence, postcode of residence and date of death. This linkage used all name information recorded on the PHDAS, and death and cancer records. Probabilistic record linkage was used to identify matching record pairs, first with the death records and then with the cancer records. The date of death determined by linkage with the NDI was used in addition to the PHDAS date of death for linkage with the ACD. For each potential matching record pair, the Australian Institute of Health and Welfare (AIHW) REMA data linkage program automatically computed a weight that reflected the likelihood of the record pair being a true match. For names, the linkage weight was based on both the similarity of the recorded names between records and the frequency of those names in the database. The weights were used to rank the matched pairs, resulting in three groups: accepted matches, non-matches and possible matches. This was followed by a clerical review of the possible matches, which were manually examined and a decision made as to their status as accepted matches. The full name record linkage was considered the gold standard in this study.

Record linkage: name code

For the name code linkage, the primary name of each registrant in the cohort, death and cancer records was truncated to the first two letters of the first name and surname, a "2x2 name code". No other name information was used for linkage; all other linkage variables were identical to those used in the full name linkage.

Rules-based deterministic record linkage using SAS® software v9·2 (SAS Institute Inc., Cary, NC, USA) was used to identify matching record pairs, first with the death records and then with the cancer records. An exact match was defined as a match on name code, date of birth, gender and state of residence. Non-exact matches were also accepted if they met pre-defined criteria (see Supplementary Tables 1 and 2, available in the online version of this article). These criteria were based on the most common mismatches between records seen during clerical review of full name linkages, along with a requirement for matching date of death or registrant's postcode between datasets. The criteria for exact and non-exact matches constituted the linkage algorithm that was coded for automatic application. The clerical review process for this linkage was thus automated due to the limited information provided by the name code.

Statistical analyses

Record linkages using the name code and full name cohort records were independently performed and the sensitivity, specificity, and positive (PPV) and negative predictive value (NPV) of the name code linkage as compared with the full name linkage (gold standard) were calculated with 95% confidence intervals (CIs). Sensitivity and specificity were also computed for cohort strata, including gender (male vs female), age at PHDAS registration (<median age vs ≥median age of 27 years), era of PHDAS registration (early NDI [pre-1997] vs later NDI [1997 onward]), whether the registrant had multiple names listed on the PHDAS (yes vs no), and whether there was a death recorded on the PHDAS (yes vs no). Sensitivity was defined as the percentage of true deaths or cancers identified by the name code linkage. Specificity was defined as the percentage of those truly alive or without cancer that did not link with a death or cancer on the respective name code linkages.

Logistic regression analyses were performed to examine which person-level and registry characteristics were associated with false-negative or false-positive linkages. The cohort characteristics examined were gender, age (tertiles) and whether the registrant had multiple names recorded on the PHDAS. When examining factors associated with missed death linkages, the year of death was also considered. Year of registration on the PHDAS was examined for false deaths and cancers and the results of the death record linkage were examined to see if they predicted the cancer linkage results. Odds ratios (ORs) and Fisher's exact 95% CIs were computed; if one or more strata had zero observations then Cornfield 95% CIs were estimated for those strata. Multivariable logistic regression was performed if more than one variable reached statistical significance ($p<0.05$) and no strata had zero observations.

Risk of death and of incident cancer for the OST cohort relative to the general population was estimated and compared for the full name and name code record linkage methods. Risk was estimated by the standardised mortality ratio (SMR) and the standardised incidence ratio (SIR), the ratios of the observed and the expected numbers of deaths and cancers respectively. The expected numbers of deaths and incident cancers were calculated by multiplying person-years at risk by Australian five-year

age-, sex-, state-, and calendar year-specific population mortality and cancer incidence rates respectively. For SMRs, person-years of follow-up accumulated from the date of first registration for OST and terminated at the date of death, age 80, or 31 December 2007 – whichever occurred first. For SIRs, person-years of follow-up accumulated from the date of first registration for OST and terminated at the date of first cancer diagnosis, death, age 80, or 31 December 2007 – whichever occurred first. The numbers of deaths and incident cancers were assumed to conform to a Poisson distribution. Overall and cause-specific mortality, and overall and site-specific cancer risk, were calculated for each record linkage method for the entire cohort. An approximation to the Poisson distribution was used to calculate 95% CIs for deaths/cancers with more than ten expected cases while exact CIs were used for deaths/cancers with fewer than ten expected cases.

Analyses were performed using SAS v9·2 (SAS Institute Inc., Cary, NC, USA). Person-years were calculated using the *%stratify* macro.[18]

The study was reviewed and approved by all relevant ethics committees (n=6) and the requirement for informed consent was waived because of the impracticality of obtaining consent from such a large number of participants. All linkages were performed by the AIHW and the researchers received only de-identified data.

Results

Record linkage accuracy

In the cohort of 45,419 OST registrants, 4,286 deaths and 979 first- or higher-order primary cancers were identified by full name record linkage. The name code linkage identified 2,799 of these deaths, resulting in a sensitivity of 65.31% (95%CI 63.88–66.73%; Supplementary Table 3). There were 39 falsely linked deaths, giving a specificity of 99.91% (95%CI 99.87–99.93%). The PPV and NPV were 98.63% (95%CI 98.13–99.02%) and 96.51% (95%CI 96.33–96.68%) respectively.

The sensitivity of the name code cancer record linkage was 76.81% (95%CI 74.04–79.42%), with 227 cancers missed (Supplementary Table 4). Specificity was 99.89% (95%CI 99.86–99.92%) with 48 falsely linked cancers. The PPV and NPV were 94.00% (95%CI 92.12–95.54%) and 99.49% (95%CI 99.42–99.56%), respectively.

The sensitivity and specificity of the death linkage was significantly lower for individuals registered with the PHDAS during the early NDI era (<1997) compared to the later NDI era (≥1997; 62.98%, 95%CI 61.28–64.65% vs 72.38%, 95%CI 69.59–75.06% and 99.85%, 95%CI 99.79–99.90% vs 99.95%, 95%CI 99.91–99.98% respectively; Figure 1). Similarly for the cancer linkage, specificity was significantly lower for those registered with the PHDAS during the early NDI era than during the later NDI era (99.84%, 95%CI 99.77–99.89% vs 99.94%, 95%CI 99.90–99.97% respectively); there was no difference in sensitivity (76.32%, 95%CI 73.04–79.39% vs 78.16%, 95%CI 72.65–83.02%; Figure 2).

Missed death record linkages

In bivariable models, false negative linkages or missed links were more common in females compared to males (OR 1.31, 95%CI 1.13–1.50) and in those with no date of death on the cohort registry (OR 1.79, 95%CI 1.48–2.17; Table 1). False negatives were less common during the later NDI era than the early NDI era (OR 0.53, 95%CI 0.46–0.60). These associations remained significant in a multivariable model.

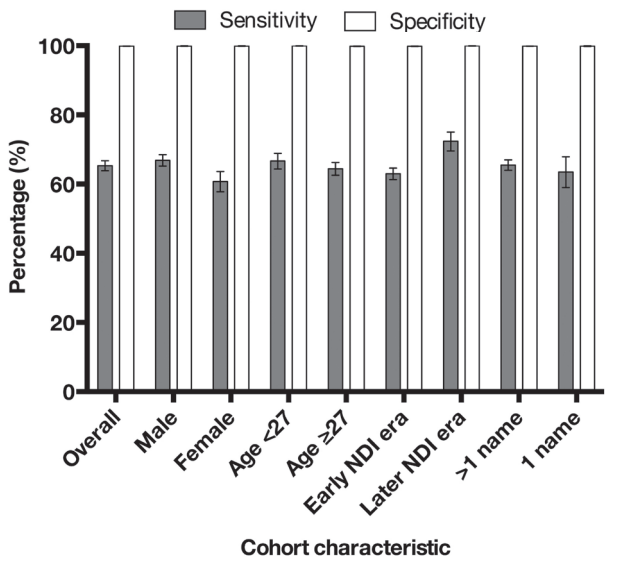
False death record linkages

NDI era was the only characteristic associated with falsely linked deaths; such links were less common in cohort members registered with the PHDAS during the later NDI era compared to the early era (OR 0.34, 95%CI 0.17–0.68).

Missed cancer record linkages

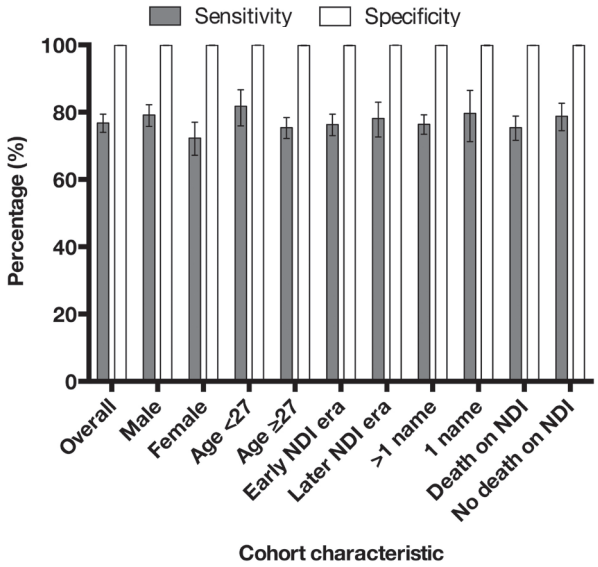
Missed cancers were more common in females than males (OR 1.45, 95%CI 1.07–1.97) and for individuals with a missed death link (OR 6.81, 95%CI 4.35–10.7). Cohort members with a linked death record were less likely to have a missed cancer compared to

Figure 1: Accuracy of death record linkage using name code compared to full name records.



NDI, National Death Index
Early NDI era: <1997
Later NDI era: ≥1997

Figure 2: Accuracy of cancer record linkage using name code compared to full name records.



NDI, National Death Index
Early NDI era: <1997
Later NDI era: ≥1997

those who were correctly identified as alive (OR 0.01, 95%CI 0.00–0.06). A zero cell in one stratum precluded multivariable analysis.

False cancer record linkages

A false death ascribed after the death linkage was positively associated with a false cancer linkage (OR 56.7, 95%CI 6.39–234), as was registrant's age in the oldest tertile relative to the youngest tertile (≥32 vs <23 years; OR 3.78, 95%CI 1.43–10.0). False cancer linkages were less likely for those registered with the PHDAS during the later NDI era compared to the early era (OR 0.35, 95%CI 0.19–0.66). A zero cell in one stratum precluded multivariable analysis.

Impact of name code linkage on risk of death

OST registrants were at an increased risk of death relative to their age and sex peers in the population regardless of whether full name or name code linkage was used. For every cause of death examined, the SMR was significantly lower using name code linkage than full name linkage (Figure 2, Supplementary Table 5). For example, the excess risk of drug-related death increased 30-fold among OST registrants using full name linkage and dropped to 20-fold using name code linkage.

Impact of name code linkage on risk of cancer

Linkage using full name records found an overall excess risk of cancer (SIR 1.16, 95%CI 1.08–1.24) whereas name code linkage found an overall reduced cancer risk (SIR 0.92, 95%CI 0.85–0.99; Figure 3, Supplementary Table 6). The increased risk of cancers of the tonsil and vulva and non-Hodgkin lymphoma was statistically significant using full name linkage but not statistically significant using name code linkage. Conversely, for mouth cancer the increased risk was statistically significant for name code linkage but not for full name linkage. Risk of cancers of the liver, pancreas, unknown primary site, larynx, lung, and cervix and Kaposi sarcoma were statistically significantly increased using both full and name code linkages, but the magnitude of the risk using name code linkage was smaller. Similarly, there was a significantly reduced risk of cancers of the colorectum, breast, brain, thyroid, and prostate and melanoma using both full and name code linkage, with lower risk estimates when the name code linkage was used. These findings are driven by a net missing number of cancers and deaths, the latter of which increases the person-years of observation and thus leads to an increase in the expected numbers of events and lower standardised incidence ratios.

Table 1: Associations between cohort and registry characteristics and the accuracy of record linkage.

Characteristic	Missed death (false-negative)		False death (false-positive)		Missed cancer (false-negative)		False cancer (false-positive)	
	OR (95% CI)	*P*-value*	OR (95% CI)	*P*-value*	OR (95% CI)	*P*-value*	OR (95% CI)	*P*-value*
Sex								
Male	1.00	<0.01	1.00	0.60	1.00	0.02	1.00	0.52
Female	1.31 (1.13–1.50)		1.91 (0.63–2.27)		1.45 (1.07–1.97)		0.81 (0.44–1.52)	
Age[a]								
Lower tertile	1.00	0.11	1.00	0.20	1.00	0.92	1.00	0.02
Middle tertile	0.89 (0.76–1.03)		1.53 (0.61–3.86)		1.08 (0.74–1.58)		2.21 (0.84–5.82)	
Upper tertile	0.83 (0.70–0.99)		2.29 (0.89–5.91)		1.08 (0.70–1.65)		3.78 (1.43–10.0)	
Number of names on cohort registry								
More than one	1.00	0.39	1.00	0.72	1.00	0.44	1.00	1.00
One	0.92 (0.75–1.12)		1.19 (0.47–3.05)		1.21 (0.75–1.94)		1.00 (0.45–2.23)	
Death record on cohort registry								
Yes	1.00	<0.01						
No	1.79 (1.48–2.17)							
Year of registration on cohort registry								
<1997			1.00	<0.01	1.00	0.55	1.00	<0.01
≥1997			0.34 (0.17–0.68)		0.90 (0.64–1.27)		0.35 (0.19–0.66)	
Year of registration on death registry								
<1997	1.00	<0.01						
≥1997	0.53 (0.46–0.60)							
Alive after deaths linkage								
No					1.00	0.21	1.00	0.57
Yes					1.21 (0.90–1.65)		1.40 (0.44–4.52)	
Comparison of full name vs name code death record linkage								
True alive					1.00	<0.01	1.00	<0.01
Missed death					6.81 (4.35–10.7)		2.12 (0.42–6.66)	
True death					0.01 (0.00–0.06)		0.00 (0.00–1.47)	
False death					0.00 (0.00–3.93)		56.7 (6.39–234)	

OR, odds ratio; CI, confidence interval.

**P-value for heterogeneity*

a: Different age definitions were required for each outcome. Namely, missed deaths = age at death; false deaths = age at cohort registration; missed cancers = age at cancer diagnosis; false cancers = age at cohort registration.

Discussion

Our study assessed the effect of 2x2 name code compared to full name linkage on two common and important outcome measures, the risk of death and the risk of cancer relative to the age-, sex- and year-matched general population. The risks of death and of cancer were significantly underestimated when name code linkage was used. The estimated risks for most individual cancer sites were lower using name code compared to full name linkage. These results suggest that missed links, or false-negatives, play an important role in the validity of name code record linkage studies and may lead investigators to underestimate true associations. As our results show, the linkage sensitivity may be so low as to invert the measures of association. The consequences of such confidentialisation of health records should therefore be weighed against their future research potential.

In our large cohort of predominantly young adult Australians, record linkage with national death and cancer records using a 2x2 name code had low sensitivity but very high specificity and PPV when compared to full name record linkage. Several factors were associated with name code record linkage accuracy, including the sex and age of the cohort members, and the extent and accuracy of the information available for linkage, such as the completeness of date of birth, and the availability of date of death for cancer linkage.

Prior studies have investigated the use of restricted name or other person-level data for record linkage with mortality records, using manual review or knowledge of vital status to determine linkage accuracy. None of these studies compared the performance of these linkage strategies to full name linkage, preventing direct comparison with our findings. Similarly, the relative accuracy of simple deterministic record linkage using a statistical linkage key, such as the SLK-581, which incorporates five letters of the name (three letters of the family name and two letters of the given name), the eight digit date of birth, and sex of the registrant, is uncertain. However, the use of an SLK in addition to person-level data items such as area of usual residence and date of health care episode has been shown, using a stepwise deterministic matching algorithm, to achieve 0.5% false positive links between aged care and residential/community care databases.[19] This approach also achieved a PPV of 97% and a sensitivity of 95% for linkage between hospitalisation and residential aged care records.[20]

A previous study found that the use of alias names may affect the quality of name data held on the various registers for these individuals.[21] Kariminia et al. (2005) also highlighted such data quality issues as an obstacle to optimal linkage accuracy when linking a prisoner population to the NDI.[16] However, as we found no association between the presence of multiple names on the PHDAS and linkage accuracy, this does not explain the high false-negative rate we observed. We performed sequential record linkage, with the results of the death linkage carried forward to the cancer linkage. We observed a higher sensitivity for the linkage with cancer records compared to linkage with death records. We hypothesise that this was due to greater accuracy and completeness of the cancer registry records due to the receipt of multiple notifications about an individual, compared to a single notification for a death register. Furthermore, date of death data is incomplete on the PHDAS and postcode data incomplete on the NDI.

Our multivariable analysis provided evidence of cohort and administrative health data characteristics that may significantly influence the rates of false-positive and false-negative linkages. Firstly, we observed increased odds

Figure 3: Cause-specific risk of death among NSW opioid substitution therapy registrants (1985–2007) by record linkage method.

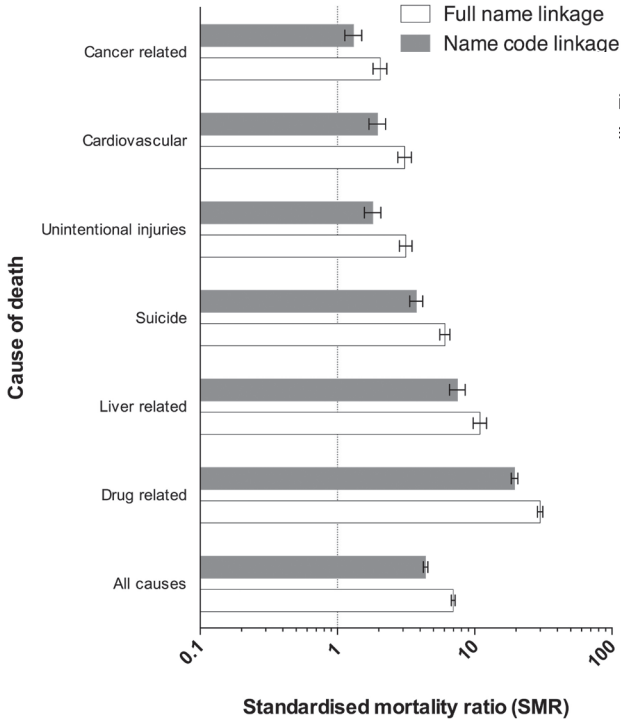
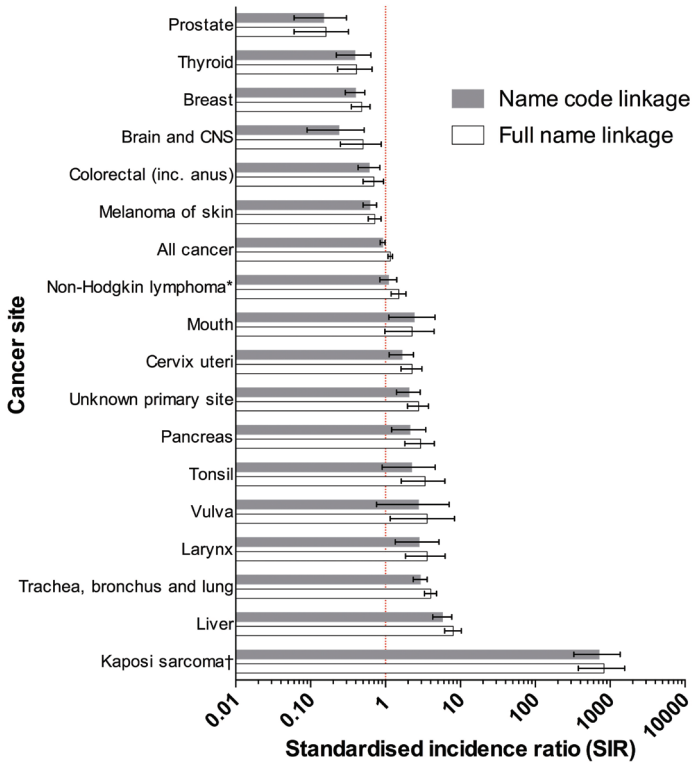


Figure 4: Site-specific risk of cancer among NSW opioid substitution therapy registrants (1985–2007) by record linkage method.



of missed death and cancer linked records for females compared to males, possibly because of the use of maiden and married surnames on different registries. Secondly, missed links were more common when there were less data with which to identify links. For example, missed death links were more likely when there was no date of death on the PHDAS and for deaths occurring during the early NDI era, where only the year of birth was recorded by the NDI (i.e. not specific dates).

As was the case in prior studies, we observed a low rate of false-positive links. Registration on the PHDAS cohort from 1997, a surrogate marker for the later NDI era, was associated with less false death and less false cancer links. A false death on the NDI linkage strongly predicted a false cancer link, likely due to the carriage of erroneous death data into the subsequent linkage. Lastly, older age was predictive of false cancer links, possibly due to a larger pool of potential cancer linkages in this age group, although this relationship was not observed for false death linkages.

The results of this study provide robust evidence for investigators performing record linkage with name coded registries and the databases utilised in this study. Record linkage is recognised as an efficient approach in observational cohort studies and its use is increasing. In Australia, privacy concerns have resulted in name codes being used for some national health registries, for example HIV/AIDS, bone marrow transplantation and cystic fibrosis. Our study provides insight into the accuracy of linkage based on records confidentialised by use of a 2x2 name code. It is likely that some of the registry characteristics explored here, such as registrant's age, gender, and data completeness, are externally valid.

Limitations

This study has several limitations. Firstly, although registrant's full name was used as the gold standard for ascertaining cancers and deaths, there remains scope for missed and false links to occur with this level of information and the estimates of sensitivity and specificity should be interpreted accordingly. A more accurate approach, linkage using a unique personal identification number only,[22-24] directly addresses privacy concerns as no personally identifying information is required, however this is not available in Australia. Secondly, one of the administrative datasets (the NDI) included an era when date of birth data was incomplete; this unique characteristic may limit the generalisability of our results for that period of the study. Also, in geographic terms, the OST registrants are a relatively transient population. This may have reduced the likelihood of a record match on the basis of postcode. Furthermore, the PHDAS did not actively seek to ascertain vital status and cohorts with more complete date of death information may have superior linkage sensitivity. Thirdly, the PHDAS is an administrative register that was not designed for epidemiological purposes. Nevertheless, it is a 'live' register that tracks registrants as they enter and exit the OST program, enabling doctors to identify and correct any data entry error and incompleteness.

The generalisability of our findings may also be affected by our use of the date of death and postcode in identifying imperfectly matched name code pairs – a method that goes some way towards objectively replicating the manual clerical review step that is possible with full names. Therefore, we cannot directly infer the linkage sensitivity and specificity of 2x2 name code linkage that does not incorporate such rules.

Finally, the stringency of linkage criteria may be adjusted in order to leverage sensitivity at the risk of reducing specificity, and vice versa. Our name code linkage used a conservative algorithm to optimise specificity, thereby minimising the likelihood of overestimating cancer incidence in this population. Researchers should consider the effect of the linkage algorithm on linkage sensitivity and specificity when working with name code records.

Conclusions

There is an increasing move towards the confidentialisation of records prior to linkage in Australia and internationally. This study demonstrates that linkage using name code records is likely to achieve conservative estimates of death and cancer risk compared to full name record linkage. Importantly, the strength of associations with outcomes such as death and cancer may be significantly underestimated, threatening scientific validity.

Acknowledgements

We thank the NSW Ministry of Health for providing the Pharmaceutical Drugs of Addiction System data, and Pia Salmelainen of the Pharmaceutical Services Branch, NSW Health, for her assistance with data extraction and interpretation. We are also grateful to the staff of the state and territory cancer registries for the use of their data. We thank the Australian Institute of Health and Welfare for conducting the data linkages.

Funding

Australian National Health and Medical Research Council (NHMRC; ID630531) and the Faculty of Medicine, University of New South Wales. Also supported by Fellowships from the NHMRC Council (MTvL ID1012141; LD ID510279 and ID1041742; AEG ID568819; CMV ID1023159) and the Cancer Institute New South Wales (CMV ID10/CDF/2-42). The National Drug and Alcohol Research Centre at the University of NSW is supported by funding from the Australian Government under the Substance Misuse Prevention and Service Improvements Grants Fund. The funding bodies played no role in the study design or conduct, data collection, analysis or interpretation of data in the writing of the article or the decision to submit the article for publication.

References

1. Lind M, Bounias I, Olsson M, Gudbjörnsdottir S, Svensson A-M, Rosengren A. Glycaemic control and incidence of heart failure in 20 985 patients with type 1 diabetes: An observational study. *Lancet*. 2011;378(9786):140-6.
2. Grulich AE, Li Y, McDonald A, Correll PK, Law MG, Kaldor JM. Rates of non-AIDS-defining cancers in people with HIV infection before and after AIDS diagnosis. *AIDS*. 2002;16(8):1155-61.
3. Olver I. Linking data to improve health outcomes. *Med J Aust*. 2014;200(7):368-9.
4. Mathews JD, Forsythe AV, Brady Z, Butler MW, Goergen SK, Byrnes GB, et al. Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: Data linkage study of 11 million Australians. *BMJ*. 2013;346:f2360.
5. Milne E, Laurvick CL, Blair E, Bower C, de Klerk N. Fetal growth and acute childhood leukemia: Looking beyond birth weight. *Am J Epidemiol*. 2007;166(2):151-9.
6. Amin J, Law MG, Bartlett M, Kaldor JM, Dore GJ. Causes of death after diagnosis of hepatitis B or hepatitis C infection: A large community-based linkage study. *Lancet*. 2006;368(9539):938-45.
7. Engels EA, Pfeiffer RM, Fraumeni JF Jr, Kasiske BL, Israni AK, Snyder JJ, et al. Spectrum of cancer risk among US solid organ transplant recipients. *JAMA*. 2011;306(17):1891-901.
8. Centre for Health Record Linkage. *Quality Assurance* [Internet]. North Sydney (AUST): CHRL; 2012 [cited 2014 Apr 22]. Available from: http://www.cherel.org.au/media/24160/qa_report_2012.pdf

9. van Leeuwen MT, Vajdic CM, Middleton MG, McDonald AM, Law M, Kaldor JM, et al. Continuing declines in some but not all HIV-associated cancers in Australia after widespread use of antiretroviral therapy. *AIDS*. 2009;23(16):2183-90.
10. Bell SC, Bye PT, Cooper PJ, Martin AJ, McKay KO, Robinson PJ, et al. Cystic fibrosis in Australia, 2009: Results from a data registry. *Med J Aust*. 2011;195(7):396-400.
11. Newgard C. Validation of probabilistic linkage to match de-identified ambulance records to a state trauma registry. *Acad Emerg Med*. 2006;13(1):69-75.
12. Grulich AE, Wan X, Coates M, Day P, Kaldor JM. Validation of a non-personally identifying method of linking cancer and acquired immune deficiency syndrome register data. *J Epidemiol Biostat*. 1996;1(4):207-12.
13. Fonseca MG, Coeli CM, de Fatima de Araujo Lucena F, Veloso VG, Carvalho MS. Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database. *Cad Saude Publica*. 2010;26(7):1431-8.
14. Nakhaee F, McDonald A, Black D, Law M. A feasible method for linkage studies avoiding clerical review: linkage of the national HIV/AIDS surveillance databases with the National Death Index in Australia. *Aust N Z J Public Health*. 2007;31(4):308-12.
15. Beauchamp A, Tonkin A, Kelsall H, Sundararajan V, English D, Sundaresan L, et al. Validation of de-identified record linkage to ascertain hospital admissions in a cohort study. *BMC Med Res Method*. 2011;11(1):42.
16. Kariminia A, Butler T, Corben S, Kaldor J, Levy M, Law M. Mortality among prisoners: How accurate is the Australian National Death Index? *Aust N Z J Public Health*. 2005;29(6):572-5.
17. New South Wales Health. *Opioid Treatment Program: Clinical Guidelines for Methadone and Buprenorphine Treatment*. Sydney (AUST): NSW Health; 2006.
18. Rostgaard K. Methods for stratification of person-time and events - a prerequisite for Poisson regression and SIR estimation. *Epidemiol Perspect Innov*. 2008;5(1):7.
19. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: The experience of the PIAC cohort study. *BMC Health Services Res*. 2010;10(1):41.
20. Australian Institute of Health and Welfare. *Movement Between Hospital and Residential Aged Care 2008–09*. Canberra (AUST): AIHW; 2013.
21. Larney S, Burns L. Evaluating health outcomes of criminal justice populations using record linkage: The importance of aliases. *Eval Rev*. 2011;35(2):118-28.
22. Chang ET, Smedby KE, Hjalgrim H, Glimelius B, Adami HO. Reliability of self-reported family history of cancer in a large case-control study of lymphoma. *J Natl Cancer Inst*. 2006;98(1):61-8.
23. Luo J, Ye W, Zendehdel K, Adami J, Adami HO, Boffetta P, et al. Oral use of Swedish moist snuff (snus) and risk for cancer of the mouth, lung, and pancreas in male construction workers: A retrospective cohort study. *Lancet*. 2007;369(9578):2015-20.
24. Sorensen GV, Ganz PA, Cole SW, Pedersen LA, Sorensen HT, Cronin-Fenton DP, et al. Use of beta-blockers, angiotensin-converting enzyme inhibitors, angiotensin II receptor blockers, and risk of breast cancer recurrence: A Danish nationwide prospective cohort study. *J Clin Oncol*. 2013;31(18):2265-72.

Supporting Information

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

Supplementary Table 1: Algorithm used for full name record linkage.

Supplementary Table 2: Algorithm used for name code record linkage.

Supplementary Table 3: Accuracy of death record linkage using name code compared to full name records.

Supplementary Table 4: Accuracy of cancer record linkage using name code compared to full name records.

Supplementary Table 5: Risk of death among NSW opioid substitution therapy registrants 1985–2007, by linkage method.

Supplementary Table 6: Risk of cancer among NSW opioid substitution therapy registrants 1985 –2007, by linkage method.