Maximising data value and avoiding data waste: a validation study in stroke research

Monique F Kilkenny[1,2], Joosup Kim[1,2], Nadine E Andrew[1,3], Vijaya Sundararajan[4], Amanda G Thrift[1], Judith M Katzenellenbogen[5,6], Felicity Flack[7], Melina Gattellari[8], James H Boyd[7], Phil Anderson[9,10], Natasha Lannin[4], Mark Sipthorp[11], Ying Chen[11], Trisha Johnston[12], Craig S Anderson[13,14], Sandy Middleton[15], Geoffrey A Donnan[2], Dominique A Cadilhac[1,2]

The known: Recent advances in digital infrastructure in Australia allow linkage of administrative and clinical datasets.

The new: This is the first time that data for patients with stroke or transient ischaemic attack from a national prospective clinical quality registry have been merged with state and national health data in Australia. A comprehensive approach to linkage quality control was undertaken.

The implications: Analyses of enriched, linked data will allow contacts with the health system before and after stroke events to be explored and enable more accurate risk adjustment of epidemiological and health care models (eg, for comorbid conditions).

Abstract

Objectives: To determine the feasibility of linking data from the Australian Stroke Clinical Registry (AuSCR), the National Death Index (NDI), and state-managed databases for hospital admissions and emergency presentations; to evaluate data completeness and concordance between datasets for common variables.

Design, setting, participants: Cohort design; probabilistic/deterministic data linkage of merged records for patients treated in hospital for stroke or transient ischaemic attack from New South Wales, Queensland, Victoria, and Western Australia.

Main outcome measures: Descriptive statistics for data matching success; concordance of demographic variables common to linked databases; sensitivity and specificity of AuSCR in-hospital death data for predicting NDI registrations.

Results: Data for 16 214 patients registered in the AuSCR during 2009–2013 were linked with one or more state datasets: 15 482 matches (95%) with hospital admissions data, and 12 902 matches (80%) with emergency department presentations data were made. Concordance of AuSCR and hospital admissions data exceeded 99% for sex, age, in-hospital death (each $\kappa = 0.99$), and Indigenous status ($\kappa = 0.83$). Of 1498 registrants identified in the AuSCR as dying in hospital, 1440 (96%) were also recorded by the NDI as dying in hospital. In-hospital death in AuSCR data had 98.7% sensitivity and 99.6% specificity for predicting in-hospital death in the NDI.

Conclusion: We report the first linkage of data from an Australian national clinical quality disease registry with routinely collected data from several national and state government health datasets. Data linkage enriches the clinical registry dataset and provides additional information beyond that for the acute care setting and quality of life at follow-up, allowing clinical outcomes for people with stroke (mortality and hospital contacts) to be more comprehensively assessed.

Several government departments share responsibility for providing health and community services in Australia, and each routinely collects data for different administrative purposes.[1,2] The federal government manages information related to births, deaths and marriages and Medicare-funded health services; state governments manage data for admissions and emergency presentations to hospitals. Clinical quality registries not managed by governments, such as the Australian Stroke Clinical Registry (AuSCR),[3] include data for more refined measures of quality of care and patient-reported outcomes (eg, health-related quality of life) than are available in government administrative datasets.

Recent advances in digital infrastructure allow linkage of administrative and clinical datasets in Australia,[4,5] but the quality of linkage has rarely been formally reported. In 2011, a 4-year National Health and Medical Research Council (NHMRC) Partnerships for Better Health grant was awarded to Dominique Cadilhac and a consortium of researchers, clinicians, and government and non-government organisations — Stroke123 — to expand existing programs for improving care in Australia for patients with acute stroke.[6] A major goal of Stroke123 was to establish a high quality, integrated, national stroke dataset for assessing the feasibility of linking a clinical quality disease registry (the AuSCR) with administrative health datasets from several states and territories.

Linkage of registry and administrative datasets allows the entire continuum of stroke care and long term outcomes to be investigated, including survival and quality of life (Box 1). In this article, we report our investigation of the feasibility of linking data from the AuSCR, the National Death Index, and state-managed databases for hospital admissions and emergency presentations. We assessed the proportion of AuSCR registrants for whom data were successfully linked, and the concordance between datasets of data for common variables.
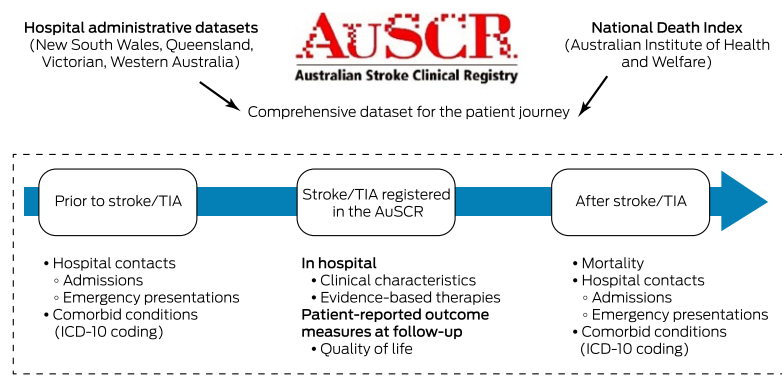
Methods

Data sources

Data from hospitals in New South Wales, Queensland, Victoria, and Western Australia that participated in the AuSCR were included in the Stroke123 data linkage project (Box 2).

The AuSCR is a prospective national stroke registry for monitoring, promoting, and improving the quality of acute stroke care in Australia (www.auscr.com.au).[3] It was designed to systematically and prospectively collect data that enable variability in clinical care processes and health outcomes for patients
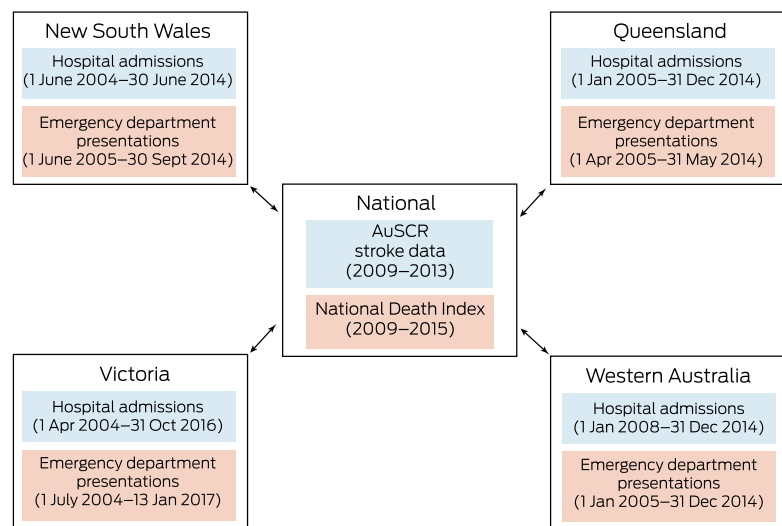
1 Linkage of registry and administrative datasets, facilitating investigation of the entire journey of patients with stroke

TIA = transient ischaemic attack. ◆

2 Data sources linked in the Stroke123 project, including years of coverage sampled

The specific linked data sources: the Victorian Admitted Episodes Data Set, NSW Admitted Patient Data Collection, Queensland Hospital Admitted Patient Data Collection, and WA Hospital Morbidity Data System for admissions; the Victorian Emergency Minimum Dataset, NSW Emergency Department Data Collection, Queensland Emergency Department Information System, and WA Emergency Department Data Collection for emergency presentations. ◆

admitted to hospital with acute stroke or transient ischaemic attack to be identified. The minimum core dataset includes demographic and clinical data; 90–180 days after admission to hospital, patients are requested to complete a follow-up questionnaire. The registry records details that facilitate routine patient follow-up, including their full name, address, sex, and date of birth. For our study, we analysed data for patients registered with the AuSCR during 2009–2013. The index stroke event for each patient was identified in the state-based datasets from the date of admission for the first-ever stroke episode registered in the AuSCR.

Fact and cause of death. The National Death Index (NDI), maintained by the Australian Institute of Health and Welfare, includes data for all deaths in Australia since 1980, including information on the date and cause of death as reported by state-based death registers.[7] Causes of death are coded according to the International Classification of Diseases, tenth revision,

Australian modification (ICD-10-AM).[8] Death data are linked by the Australian Institute of Health and Welfare with AuSCR data annually.

Hospital admissions. Each Australian state and territory collects details on all inpatient separations (discharges, transfers, deaths) from all public, private, psychiatric, and repatriation hospitals in their state or territory, including demographic data, dates of admission and discharge, marital status of the patient, urgency of admission, funding source, hospital insurance, ICD-10 diagnosis codes, and procedure codes (Australian Classification of Health Interventions[9]).

Emergency department presentations. Information on presentations to public (but not private) emergency departments collected includes demographic data, date and time of presentation, urgency of presentation, triage category, and diagnoses.

Data linkage

Australian residents do not have unique personal identifiers, necessitating the application of probabilistic or deterministic data linkage techniques (Box 3). The data linkage methods applied in the Stroke123 project have been described previously.[1] In brief, the AuSCR data manager submitted each registrant's name, date of birth, sex, address, admission and discharge dates, and other personal identifiers (Medicare number and, in some states, unit record number) to data linkage units in each of the participating states. Data linkage unit staff linked patient details and demographic data as accurately as possible with deterministic or probabilistic linkage software, and also undertook manual clerical reviews if required. A special purpose data linkage process was developed for the Stroke123 project, the first trial of this method for cross-jurisdictional linking of clinical quality registry data with data from administrative databases.[1] Each data linkage unit submitted a state-specific linkage key, matched to the AuSCR project linkage key, to the Centre for Data Linkage (CDL) at Curtin University, Perth. The CDL was established as a secure data linkage facility that facilitates linkage of jurisdictional datasets on the basis of demographic data. The CDL created data linkage keys that they communicated to data analysts to enable merging of de-identified datasets across jurisdictions.

Data management

As the states have separate departments of health and health information systems, data collection is not standardised across Australia. Consequently, not only did the variables for which data were collected and their names differ between states, the data formats (eg, numeric or alphanumeric) and options for variables (eg, marital status: de facto category included or excluded) also differed. In consultation with an expert working group (the Stroke123 data linkage subcommittee), we developed a data dictionary that detailed how consolidated variables for the merged dataset should be created from the data in the state-based datasets. The subcommittee comprised experts in data linkage processes (ie, ethics and data custodian approvals), representatives from the individual data linkage units, researchers, and

3 Australian Stroke Clinical Registry (AuSCR) registrants who were matched with data in state administrative datasets

		Dataset*			
	Total	**New South Wales**	**Queensland**	**Victoria**	**Western Australia**
Linkage method		Probabilistic	Probabilistic and deterministic	Stepwise deterministic	Probabilistic
Number in registry	16 214	4090	5616	5529	999
Number matched					
Hospital admissions	15 482[†] (95%)	3714 (91%)	5524 (98%)	5278 (95%)	971 (97%)
Emergency department presentations	12 902[‡] (80%)	3830 (94%)	4202 (75%)	3929 (71%)	941 (94%)
Hospital and emergency presentations	15 842 (98%)	3951 (97%)	5566 (99%)	5317 (96%)	998 (99%)

* Linkage of state database and Australian Stroke Clinical Registry (AuSCR) data undertaken by New South Wales Centre for Health Record Linkage (CHeReL); the Victorian Data Linkage Unit; Queensland Health; the Western Australian Data Linkage branch. † 15 482 registrants with 15 487 admissions. ‡ 12 902 registrants with 13 985 presentations. ◆

clinicians. Coding of the consolidated variables was verified independently by two data analysts (authors MFK, NEA) and any discrepancies discussed by the subcommittee during the development phase to ensure that options and data types for variables were comparable. Advice was also sought from researchers who had previously led projects involving the merging of data from multiple state datasets.[5]

Statistical analysis

All analyses were performed in Intercooled Stata/SE 14.0 for Windows (StataCorp). The agreement of data for variables common to the admissions and AuSCR datasets (age, sex, country of birth, Indigenous status, admission date, in-hospital death) was assessed with the standard Cohen κ statistic ($\kappa > 0.8$, excellent agreement; $0.6 < \kappa \leq 0.8$, good agreement).[10] As data for patients exist in the death registry dataset only if the patient has died, we determined the accuracy of these data by comparing them with AuSCR death data. We calculated the sensitivity and specificity of in-hospital death data in the AuSCR dataset for matching in-hospital death according to the death date recorded in the NDI (the gold standard). Sensitivity was calculated as the number of true positives divided by the number of people classified as dead according to the date of death recorded in the NDI, and specificity as the number of true negatives divided by the number of people classified as alive by the NDI in-hospital death variable. Positive predictive value was calculated as the number of true positives divided by the number of people classified as having died by the AuSCR, negative predictive value as the number of true negatives divided by the number of people classified as alive by the AuSCR.

Ethics approval

An overarching ethics approval for Stroke123, including approval for the data linkage substudy, was provided by the Monash University Human Research Ethics Committee (reference, CF13/1303 – 2013000641) and the MetroSouth Human Research Ethics Committee in Queensland (reference, HREC/13/QPAH/31). Overall ethics approval was required in Queensland as other components of the Stroke123 project were undertaken there. State and federal ethics approvals for data linkage were also required and provided by the Australian Institute for Health and Welfare (to link AuSCR registrants' data to NDI data; reference, EO 2013/2/16), the New South Wales Population and Health Services Research Ethics Committee (reference, HREC/14/CIPHS/66), and the Western Australia Department of Health Human Research Ethics Committee (reference, 2015/33).

The Victorian Department of Health accepted the ethics approval provided by the NSW Department of Health for linkage of data from Victoria.

Results

After removing data for 102 registrants who had opted out of submitting data to the AuSCR and data for patients with missing admission dates (two AuSCR registrants and 919 admissions), data for 16 214 AuSCR registrants were available for matching with data for 243 892 admissions in the four participating states, including 15 482 registrants (95% of AuSCR registrants) with data in both the admissions and AuSCR datasets (Box 4). A total of 732 patients in the AuSCR database (5%) could not be matched to an admission at the time of, before, or after their index stroke event (92 in Queensland, 376 in New South Wales, 28 in Western Australia, 251 in Victoria).
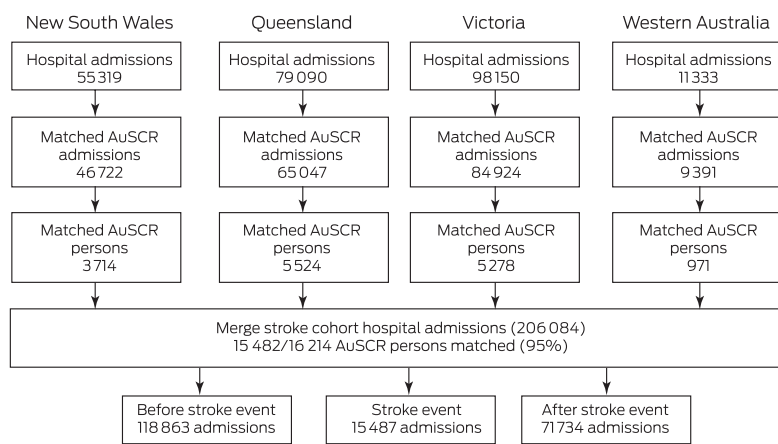
Data from the same AuSCR registrants were further matched with 90 823 emergency presentations in the four states; data for 12 902 patients were available in both datasets. There were no records of emergency presentations for 3332 patients with an AuSCR record (20%): 1600 in Victoria (29%), 1414 (25%) in Queensland, 260 (6%) in New South Wales, and 58 in Western Australia (6%) (Box 3).

In total, 15 842 registrants in AuSCR (98%) were matched to a hospital admission or an emergency presentation (or both) in one or more of the four states (Box 3). No admissions or emergency presentations were recorded for 372 patients (2%) in the AuSCR: 50 (1%) from Queensland, 139 (3%) from New South Wales, 212 (4%) from Victoria, and one (< 1%) from Western Australia. The rate of matching between AuSCR data and data for admission or emergency presentation in any of the four states was 98%.

Concordance of the AuSCR and hospital admissions datasets was excellent for the demographic variables sex ($\kappa = 0.99$), age ($\kappa = 0.99$), and Indigenous status ($\kappa = 0.83$), and also for in-hospital death ($\kappa = 0.99$) (each > 99% concordance; Box 5).

NDI records were available for 4183 of 16 214 AuSCR registrants; of 1498 registrants identified in the AuSCR as dying in hospital, 1440 were also recorded by the NDI as dying in hospital (96%). In-hospital death in AuSCR data had 98.7% sensitivity and 99.6% specificity for predicting in-hospital death in the NDI. The positive predictive value of the AuSCR in-hospital death status was 96.1%, the negative predictive value 99.9%. For 77 AuSCR

4 Matching of Australian Stroke Clinical Registry (AuSCR) registrants and patients in state hospital admissions data

New South Wales	Queensland	Victoria	Western Australia
Hospital admissions 55 319	Hospital admissions 79 090	Hospital admissions 98 150	Hospital admissions 11 333
Matched AuSCR admissions 46 722	Matched AuSCR admissions 65 047	Matched AuSCR admissions 84 924	Matched AuSCR admissions 9 391
Matched AuSCR persons 3 714	Matched AuSCR persons 5 524	Matched AuSCR persons 5 278	Matched AuSCR persons 971

Merge stroke cohort hospital admissions (206 084)
15 482/16 214 AuSCR persons matched (95%)

Before stroke event 118 863 admissions	Stroke event 15 487 admissions	After stroke event 71 734 admissions

5 Concordance of Australian Stroke Clinical Registry (AuSCR) and state administrative datasets for variables common to linked datasets

Variable	Agreement	κ (95% CI)	Variable not available
Age group*	99.89%	0.99 (0.999–0.999)	Queensland, Western Australia
Sex (male)*	99.75%	0.99 (0.995–0.995)	Queensland
Indigenous status	99.46%	0.83 (0.818–0.836)	Queensland, Victoria
Australian-born	92.42%	0.84 (0.834–0.845)	Queensland
In-hospital death	99.32%	0.96 (0.960–0.962)	—
Admission date	88.08%	—	—

CI = confidence interval. * Matching variables used for data linkage of the AuSCR and administrative datasets. ◆

for all registrants would also exist in state admissions datasets. The lower proportion of matches between the AuSCR and emergency datasets was similarly expected, and differences between states probably reflect differences in care pathways. For example, some patients with stroke or transient ischaemic attack in Queensland and Victoria are admitted directly to hospital, bypassing the emergency department. Further, data collected in private hospital emergency departments were available only for Western Australia.

The quality of AuSCR data are monitored routinely in random audits and reviews of case ascertainment. However, linkage to the hospital administrative datasets and comparison of data for variables common to both dataset types provided another opportunity for validating AuSCR data. We found that the accuracy of AuSCR data was comparable with that of a similar United States registry, the Get With The Guidelines (GWTG) stroke registry; auditors who compared GWTG data with medical records found an overall agreement of greater than 90% for a range of variables.[11]

Linkage with the NDI provided an opportunity to assess the accuracy of the mortality data in the AuSCR. Some discrepancies were found, perhaps attributable to errors of data entry at the AuSCR. Conversely, some linkages may be missed because of errors in the NDI caused by incorrect information on death certificates.[12] Other investigators, based on comprehensive follow-up of deaths in their studies, have suggested that NDI mortality data may be incomplete.[13–17] According to these studies, the NDI captured at least 88% of deaths verified by the authors, but the investigations typically examined deaths after hospital discharge, and it is likely that notifications outside of hospital admissions are less reliably recorded.

Reliable linkage of routinely collected health information systems data with AuSCR data will enhance the primary function of the registry by providing reliable information on outcomes, including survival and hospital contacts. This information will improve stroke care by guiding policy and practice. It will be of particular value in stroke care because patients interact with several health care facilities, sometimes in different states; it would not be feasible to collect these data directly from all registrants.

Clinical diagnoses recorded by the AuSCR and discharge diagnoses in hospital admissions records will be examined in future validation work. Discrepancies of clinical and principal diagnoses in the two datasets will be employed to educate health information managers who code the diseases for hospital presentations, ultimately improving the coding of stroke in Australian hospitals and ensuring that it is closely aligned with clinical diagnoses by neurologists and other medical staff. Data linkage as described in our article will also allow exploration of hospital contacts before and after stroke events in analyses of health care utilisation following stroke or transient ischaemic attack.

registrants, discordant death information was recorded in the two datasets: 19 (25%) were recorded by the NDI but not by the AuSCR as dying in hospital, while 58 (75%) were recorded by the AuSCR as dying in hospital but by the NDI as dying after discharge (21 patients) or still living 180 days after discharge (37 patients) (Box 6).

Discussion

We report the first matching of data from an Australian clinical quality registry with data from state and national population health datasets. We found that matching was reliable (98% matched with hospital admission or emergency department presentation data), that concordance of information for demographic variables in the linked datasets was excellent, and that AuSCR data on in-hospital deaths matched an NDI record of in-hospital death with high sensitivity and specificity.

The AuSCR includes data for patients admitted to hospital with stroke or transient ischaemic attack, and we expected that details

Limitations

One limitation of the AuSCR dataset is that it does not include detailed information on comorbid conditions, diagnoses, and procedures. It has recently been noted that analysing AuSCR data without information on covariates such as comorbid conditions and stroke risk factors may result in residual confounding.[18] The enriched dataset we have described (AuSCR and admissions and emergency presentations data) will increase the

6 Survival status of Australian Stroke Clinical Registry (AuSCR) registrants*

		Time of death, according to National Death Index (NDI)				
		In hospital	After discharge, within 90 days of admission	91–180 days after admission	Alive 180 days after admission[†]	Total
In-hospital death status, according to AuSCR	Died	1440	20	1	37	1498
	Alive	19	993	413	13 291	14 716
	Total	1459	1013	414	13 328	16 214

* Based on a derived AuSCR in-hospital death variable. † 1297 of 13 328 patients alive 180 days after admission had a record in the NDI. Cells with discrepant data (ie, in-hospital death recorded in one dataset only) are shaded. Calculations: sensitivity: 1440/1459 = 98.7%; specificity: (993 + 413 + 13 291)/(1013 + 414 + 13 328) = 99.6%; positive predictive value: 1440/1498 = 96.1%; negative predictive value: (993 + 413 + 13291)/14 716 = 99.9%. ◆

explanatory power of multivariable analysis regression models with risk adjustment by including data on comorbid conditions (based on ICD-10 codes associated with hospital presentations during a 5-year lookback period) in statistical models.[19]

An additional limitation of the AuSCR is that dates of death are entered by clinicians according to information available in hospital medical records. The survival status of patients transferred to other hospitals is not captured in the AuSCR.

Conclusion

Data linkage enriches the clinical registry dataset and provides additional information beyond the acute care setting and quality of life at follow-up, facilitating a more comprehensive assessment of clinical outcomes after stroke (mortality and hospital contacts). We have shown it is feasible to accurately link and merge data from a clinical quality registry with national and state government health data from multiple jurisdictions. Interrogating this dataset will provide essential evidence — including information about missed opportunities for preventing stroke and providing post-stroke support — that will improve the quality of clinical care and, ultimately, health outcomes for patients with stroke or transient ischaemic attack.

1 Andrew NE, Sundararajan V, Thrift AG, et al. Addressing the challenges of cross-jurisdictional data linkage between a national clinical quality registry and government-held health data. *Aust N Z J Public Health* 2016; 40: 436–442.

2 Australian Institute of Health and Welfare. National minimum data sets and data set specifications. http://meteor.aihw.gov.au/content/index.phtml/itemId/344846 (viewed Sept 2018).

3 Cadilhac DA, Lannin NA, Anderson CS, et al. Protocol and pilot data for establishing the Australian Stroke Clinical Registry. *Int J Stroke* 2010; 5: 217–226.

4 Olver IN. Linking data to improve health outcomes. *Med J Aust* 2014; 200: 368–369. https://www.mja.com.au/journal/2014/200/7/linking-data-improve-health-outcomes

5 Spilsbury K, Rosman D, Alan J, et al. Cross border hospital use: analysis using data linkage across four Australian states. *Med J Aust* 2015; 202: 582–586. https://www.mja.com.au/journal/2015/202/11/cross-border-hospital-use-analysis-using-data-linkage-across-four-australian

6 Cadilhac DA, Sundararajan V, Andrew N, et al. Using linked data to more comprehensively measure the quality of care for stroke: understanding the issues. *Australasian Epidemiologist* 2013; 20: 15–19.

7 Australian Institute of Health and Welfare. About National Death Index. Updated June 2018. https://www.aihw.gov.au/about-our-data/our-data-collections/national-death-index/about-national-death-index (viewed Sept 2018).

8 National Centre for Classification in Health. The International statistical classification of diseases and related health problems, tenth revision, Australian modification (ICD-10-AM). Wollongong: National Casemix and Classification Centre, University of Wollongong, 2013.

9 National Centre for Classification in Health. Australian classification of health interventions (ACHI). Sydney: NCCH, 2008.

10 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.

11 Xian Y, Fonarow GC, Reeves MJ, et al. Data quality in the American Heart Association Get With The Guidelines-Stroke (GWTG-Stroke): results from a national data validation audit. *Am Heart J* 2012; 163: 392–398, 398.e1.

12 Australian Institute of Health and Welfare. National Death Index (NDI), data quality statement. Metadata Online Registry (METeOR). Sept 2012. http://meteor.aihw.gov.au/content/index.phtml/itemId/480010 (viewed July 2018).

13 Nagle CM, Purdie DM, Webb PM, et al. Searching for cancer deaths in Australia: National Death Index vs. cancer registries. *Asian Pac J Cancer Prev* 2006; 7: 41–45.

14 Kelman C. The Australian National Death Index: an assessment of accuracy. *Aust N Z J Public Health* 2000; 24: 201–203.

15 Magliano D, Liew D, Pater H, et al. Accuracy of the Australian National Death Index: comparison with adjudicated fatal outcomes among Australian participants in the Long-term Intervention with Pravastatin in Ischaemic Disease (LIPID) study. *Aust N Z J Public Health* 2003; 27: 649–653.

16 Kariminia A, Butler T, Corben S, et al. Mortality among prisoners: how accurate is the Australian National Death Index? *Aust N Z J Public Health* 2005; 29: 572–575.

17 Powers J, Ball J, Adamson L, Dobson A. Effectiveness of the National Death Index for establishing the vital status of older women in the Australian Longitudinal Study on Women's Health. *Aust N Z J Public Health* 2000; 24: 526–528.

18 Cadilhac DA, Andrew NE, Lannin NA, et al. Quality of acute care and long-term quality of life and survival: the Australian Stroke Clinical Registry. *Stroke* 2017; 48: 1026–1032.

19 Preen DB, Holman CD, Spilsbury K, et al. Length of comorbidity lookback period affected regression model performance of administrative health data. *J Clin Epidemiol* 2006; 59: 940–946. ■

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article.