

Linking ambulance, emergency department and hospital admissions data: understanding the emergency journey

Julia L Crilly, John A O'Dwyer, Marilla A O'Dwyer, James F Lind, Julia A L Peters, Vivienne C Tippet, Marianne C Wallis, Nerolie F Bost and Gerben B Keijzers

Both the challenges and the necessity of using routinely collected data for monitoring health services are well recognised. Within Queensland Health's eHealth strategy, the use of data to design and monitor continuity of care services is a priority.¹ For patients treated in an emergency department (ED), data are collected in many different information systems, often without a unique or common identifier to facilitate tracking of the patient across these systems.

When a common identifier is not available, data linking can be used to collate patient data. As well as tracking an individual's episodes of care across time, linked data have been used to provide a more comprehensive analysis of population health needs;^{2,3} enable longitudinal analyses of trends and patterns of care;³ monitor and evaluate health care services,³⁻⁶ practices² and policies;^{3,7} describe the health status of communities;^{3,4,7} monitor the effect of interventions for particular medical conditions;⁸ and detect and avert poor outcomes.⁶

In Australia, the Western Australian data linkage system⁹ is well established and has been used to link data from a variety of databases (eg, hospital morbidity databases, cancer registries, births and death registries, mental health services, midwives' notifications).¹⁰

Despite the advantages of data linkage, progress in linking data within Australian states (other than WA), and across states, has been slow. This is largely due to factors such as concerns about privacy safeguards,^{3,11} the ready availability of data, different health datasets being collected by different levels of government⁵ or different organisations,⁴ and the lack of unique identifiers.^{12,13}

Recently, the federal government introduced legislation to support a national approach to developing individual electronic health records, facilitated by a national Unique Healthcare Identifier (UHI). While the UHI should enable better collation of data, it is as yet unclear what level of data — summary demographic or specific detailed clinical information —

ABSTRACT

Objective: To assess the accuracy of data linkage across the spectrum of emergency care in the absence of a unique patient identifier, and to use the linked data to examine service delivery outcomes in an emergency department (ED) setting.

Design: Automated data linkage and manual data linkage were compared to determine their relative accuracy. Data were extracted from three separate health information systems: ambulance, ED and hospital inpatients, then linked to provide information about the emergency journey of each patient. The linking was done manually through physical review of records and automatically using a data linking tool (Health Data Integration) developed by the CSIRO (Commonwealth Scientific and Industrial Research Organisation). Match rate and quality of the linking were compared.

Setting: 10 835 patient presentations to a large, regional teaching hospital ED over a 2-month period (August – September 2007).

Results: Comparison of the manual and automated linkage outcomes for each pair of linked datasets demonstrated a sensitivity of between 95% and 99%; a specificity of between 75% and 99%; and a positive predictive value of between 88% and 95%.

Conclusions: Our results indicate that automated linking provides a sound basis for health service analysis, even in the absence of a unique patient identifier. The use of an automated linking tool yields accurate data suitable for planning and service delivery purposes and enables the data to be linked regularly to examine service delivery outcomes.

MJA 2011; 194: S34–S37

will be available to health care professionals and researchers. In addition, there are likely to be other databases, not linked by the UHI, which may be necessary for effective research and health service planning.

Three main data linkage methods — manual, probabilistic and deterministic — have been reported in the research literature (Box 1). All three are prone to false negatives and false positives.¹⁶

The use of information technology data linkage strategies (by deterministic or probabilistic approaches), to replace conventional time-consuming manual data linkage, has the potential to inform and benchmark services on a scale not previously possible within Queensland. We report on the accuracy of manual and deterministic linking of three health information systems using 2 months of demographic, clinical and service delivery data.

METHODS

Data sources

Data for linking were sourced from three health information systems: the Queensland

Ambulance Service Electronic Ambulance Report Form (eARF), the Emergency Department Information System (EDIS), and the Hospital Based Corporate Information System (HBCIS). Box 2 shows the specific data sourced from each system. The information collected was based on previous ED research reports.^{4,15,19,20}

Data linking methods

Our standard initial data-checking processes included a brief manual "clean" of each of the three datasets to remove a few ($n = 483$, 4.2%) obvious data entry discrepancies (eg, no name, no date of birth). Two different methods of linking were then applied to the health information system data for each patient presentation made to the ED during a 2-month time frame (August and September 2007). The aim of linking was to identify episodes of care, including the patient's acute illness in the ED, plus or minus the ambulance episode of care, plus or minus the hospital admission, as provided.

The first linking method — manual linking — was performed by one of the researchers (JLC) with previous experience

1 The three main data linkage methods

Manual linking requires human labour and involves visually comparing two (or more) datasets and determining whether each individual episode or patient is the same across datasets. A manual "cut and paste" is then required to merge each matching episode from each dataset into one final complete dataset. Manual linkage is often performed on relatively small datasets, and is extremely time consuming and expensive.⁴ Although it is not perfect because of the possibility of human error and the labour costs, in the absence of automated linking software, it is a standard approach for combining datasets for subsequent analysis.^{14,15}

Probabilistic linking involves linking records in two (or more) files and is based on the probabilities of agreement and disagreement between a range of match variables.¹⁶ It is reported to be accurate and gives a high linkage yield.¹⁷

Deterministic linking involves linking records based on exact agreement of the selected match variables.¹⁶ Using a specifically constructed algorithm, deterministic linking can successfully identify valid links¹⁶ by allowing for the inclusion of unique variables (eg, date and time of admission) within a linking algorithm.¹⁸ It is a linkage method that is less often cited, but has been used in large scale linkages involving multiple databases.¹⁶

of this process. The second method was deterministic linking and involved the use of the Health Data Integration (HDI) software developed by the CSIRO (Commonwealth Scientific and Industrial Research Organisation),¹⁸ with the configuration of linking algorithms being performed by experienced CSIRO software consultants.

For both approaches, the linking variables were name, sex, and age (± 5 years). To increase the accuracy of identification, linking also involved using data related to date and time of arrival at the ED (for ambulance-ED data), and date and time of ED discharge and date and time of hospital admission (for ED-hospital admission).

The HDI software uses demographic data (eg, surname, first initial, sex) to link patient records, and caters for basic data entry errors and misspellings. It was also configured to link on date and time. In its deterministic-linking approach, it applies a set of rules to specify the combination of fields which indicate that the records refer to the same patient. The HDI linking

2 Data obtained from each health information system

Data type	Electronic Ambulance Report Form (eARF)	Emergency Department Information System (EDIS)	Hospital Based Corporate Information System (HBCIS)
Demographic	<ul style="list-style-type: none"> Name Age Sex 	<ul style="list-style-type: none"> Name Age Sex 	<ul style="list-style-type: none"> Name Age Sex
Clinical	<ul style="list-style-type: none"> Reason for transport 	<ul style="list-style-type: none"> Reason for presentation Mode of arrival Australasian triage score (ATS) Discharge destination from ED ICD-10-AM code 	<ul style="list-style-type: none"> ICD-10-AM code Discharge destination from hospital
Service delivery	<ul style="list-style-type: none"> Date and time of arrival to ED by ambulance Date and time of ambulance triage Date and time of ambulance patient offload Date and time of ambulance departure from ED 	<ul style="list-style-type: none"> Date and time of arrival Date and time seen by ED doctor Date and time of discharge from ED 	<ul style="list-style-type: none"> Date and time of hospital admission Date and time of hospital discharge

ED = emergency department. ICD-10-AM = International classification of diseases, 10th revision, Australian modification.

results were validated against the manual linking technique to determine sensitivity, specificity and positive predictive value (PPV).

Investigating health service delivery outcomes

Data linking was undertaken to investigate health service delivery outcomes over a 24-month period across three hospitals and the region's ambulance service after the opening of a new ED in the same health service district.

For this report, the initial data linking was applied to 2 months of data (1 month before and 1 month after the opening of the new ED) to test the clinical application of the linked data.

Ethics approval

Ethics approval to undertake our research was granted by the Health Service District Human Research Ethics Committee and the Queensland Ambulance Service. Approval from the Director General of Queensland Health to access and use health information for research was also sought and granted.

RESULTS

Data from the 2-month period used for linking comparisons included: 3469 ambulance records, 10 835 ED records and 3431 hospital admission records. Manual linking resulted in the ED records being linked with 3192 (92.0%) of the ambulance records and 3244 (94.5%) of the hospital admission records. Deterministic linking with the HDI software resulted in the ED records being linked with 3049 (87.9%) of the ambulance records and 3260 (95.0%) of the hospital admission records.

Validation of the HDI linking against manual linking revealed some false positives (ie, the HDI linked the data but the manual approach did not): $n = 1$ and $n = 55$ for the ambulance-ED and ED-hospital admission datasets, respectively; and false negatives (ie, the HDI did not link the data but the manual approach did): $n = 144$ and $n = 39$ for the ambulance-ED and ED-hospital admission datasets, respectively).

The sensitivity, specificity and PPV for the HDI-linked data compared with the manually linked data were as follows: the ambulance-ED linkage had a sensitivity of 95.5%, a specificity of 99.6% and a PPV of 87.9%;

the ED–hospital admissions linkage had a sensitivity of 99.0%, specificity of 74.9% and PPV of 95.0%.

Critical to assessing the utility of these methods for continued surveillance were the comparative times taken to perform the linking: the HDI linking took only 5 minutes compared with 200 hours for the manual approach, although the HDI linking involved an initial setup time of about 80 hours for customising and checking the HDI-linking algorithm. However, once this initial setup had been performed, the tool was used to link the data quickly on an ongoing basis.

Interim analysis of 2 months of clinical data was undertaken with both the manually linked and the HDI-linked datasets. This showed that admission rates for patients presenting to the ED were similar in both datasets.

DISCUSSION

Our findings indicate that, compared with manual linking, deterministic linking of data from three sources using the HDI software has a high sensitivity and slightly lower specificity, and is of sufficient accuracy to provide a linked dataset for use in future emergency health care research.

Differences between sensitivity and specificity are not unique to our study. According to Chu, determining sensitivity and specificity (using results from true positives, true negatives, false positives and false negatives) usually involves some degree of trade off.²¹ False positives affect specificity while false negatives affect sensitivity.²¹

More true negatives were present in the linkage between the ambulance and ED datasets (than in the ED and hospital admissions datasets). In-depth interrogation of these cases revealed that insufficient data were available for the link to occur (eg, first name only, standard date-of-birth entry used for patients whose date of birth was missing). This is the nature of the environment where these data are captured (ie, a pre-hospital and ED environment); the focus is on saving the lives of people who may not be able to provide information, nor have identification on them that would allow the required data entry fields to be completed. The smaller number of true negatives between the ED and hospital admissions datasets (compared with the ambulance–ED link) likely reflect opportunities for data capturing procedures over longer periods of time within the hospital setting.

False positives also affected specificity in our study. More false positives were noted in the ED–admissions link compared with the ambulance–ED link. In-depth interrogation of these cases revealed that the HDI linking was correct, and human error accounted for cases not being linked by the manual process, when in fact they should have been. False negatives also affected linkage sensitivity in our study. In-depth interrogation of these cases revealed that the HDI was correct in not linking the cases, due to the specifically designed algorithm rules.

Previous studies using linked data for health care research have not always provided an accompanying report on the accuracy (sensitivity and specificity) of the linked dataset. Of the research identified that used probabilistic data linkage, sensitivity rates have ranged from 88.4% to 94.6%^{13,17} and specificity rates from 99.7% to 99.8%.^{13,17} Of the identified research that used deterministic data linkage, sensitivity rates have ranged from 90% to 92%²² and specificity rates of 100%²² have been reported. Our sensitivity rates (96% and 99%) were higher than those in other reports; however, our specificity rates (75% and 99%) were comparable or lower, dependent on the datasets linked.

Exact numbers or linkage standards for defining acceptable linkage accuracy were not able to be identified from the research literature. Some authors consider that linkage accuracy is acceptable if statistically valid conclusions can be drawn.²³ Thus, not attaining 100% linkage accuracy does not prevent the use of linked data for research purposes.²³

Other studies using deterministic and probabilistic approaches to link data from various datasets have achieved varying linkage rates, ranging from 85% to 99% with probabilistic approaches^{4,17} and from 73% to 88% with deterministic approaches.^{16,22} Our findings (with deterministically linked data) are comparable, with a linkage rate of 88% for ambulance–ED data and 95% for ED–hospital admission data. The study closest to our objectives, which also involved linking ambulance and ED data, used a probabilistic approach and reported a preliminary linkage rate of 85%.⁴ Another study comparing probabilistic and deterministic record linkage of seven different data sources in the United States for a statewide trauma registry reported similar matching results.²⁴ Either approach appears to be useful when linking multiple datasets.

Clearly, the quantity and quality of data within datasets can affect the linkage rate. Furthermore, the types of datasets that have been linked vary widely. Examples include registry data from multiple hospitals and a social security death register;²² general practitioner data, hospital admissions data and social services data;¹⁷ and aged care assessment data, residential aged care data, data on extended aged care at home, home and community care data, veterans' home care, and national death index data.¹⁶ It is difficult to make further comparisons and inferences between linkage approaches and linkage rate yields, given the variety and nature of the datasets that have been linked, the different health systems from which data were drawn, and the different linking approaches used.

In our study, not only was the deterministic approach to data linking accurate, but the HDI software was also flexible enough to allow for the inclusion of novel fields in the linking algorithm; in this case, the use of time-of-event data (eg, date and time of arrival). This variation on the usual case (name) matching process was useful in linking separate ambulance, ED and hospital data to gain an accurate record of a patient's episode of care. An interim analysis conducted with each of the linking approaches revealed similar results. Furthermore, our results for hospital admission rate were similar to those given in a national report from the Australian Institute of Health and Welfare.²⁵

Linkage between units of service (eg, ambulance and ED) must be based on a logically supported and methodologically sound approach,⁸ to allow the critical evaluation of some or all aspects, of a patient's episode of care. Within specific practice areas relying on linked data, clinician input into service evaluation research is imperative so that the findings reported are clinically relevant.²⁶ To plan, manage and evaluate all levels of care, data linkage must become part of standard practice.⁵

Research arising from data linkage systems within WA has been able to successfully influence policy decisions as well as clinical practice.⁷ We undertook our data linkage project to investigate the impact on ambulance, ED, hospital and patient outcomes of opening an additional ED within a Queensland region. It is a Queensland Health priority to expand hospital and related services to meet growing community need.¹ Within Queensland alone, there are currently at least six hospitals undergoing

redevelopment or expansion, or building additional emergency services. Our data linkage project is therefore of practical and clinical importance for service planning and research on non-disparate outcomes.

The limitations of our study include the use of health information system data. As secondary data, their reliability and validity may be questioned. However, they are frequently used for research.²⁷ The absence of a unique identifier is often reported as a limiting factor,^{12,13} and the inclusion of unique identifiers is a recommendation for enhancing linkage rates across systems.^{12,13} Until unique patient identifiers become available, accurate data linking is required to undertake large-scale research investigating the patient journey.

CONCLUSION

Our study has established the benefit of automated deterministic data linkage in Queensland. Our method has generated efficient and accurate linking results and a correct patient journey record in much less time than the manual method. Application of this tool could facilitate timely routine performance monitoring and longer range benchmarking. The deterministic linkage of the three health information system datasets is being used to inform a 24-month pre-post study to examine the implications of opening an additional ED within a busy regional health service district.

ACKNOWLEDGEMENTS

We wish to thank David Hansen from the Australian e-Health Research Centre for critical review of the manuscript. We also wish to acknowledge funding received from the Queensland Ambulance Service to undertake the research described in this article.

COMPETING INTERESTS

None identified.

AUTHOR DETAILS

Julia L Crilly, BNurs, MN(Hons), PhD, Nurse Researcher, Associate Professor¹

John A O'Dwyer, BCompScInEng(Hons), Technical Consultant²

Marilla A O'Dwyer, BEng(Hons), Senior Consultant²

James F Lind, BS BM, BMedSci, FACEM, Associate Professor, Executive Director of Emergency Medicine Training¹

Julia AL Peters, MB BS, Registrar³

Vivienne C Tippett, BA, GradDipPsych, MPH(BiostatEpi), Director⁴

Marianne C Wallis, RN, BSc(Hons), PhD, Chair, Clinical Nursing Research, Professor⁵

Nerolie F Bost, BNurs, MN, Research Nurse Officer¹

Gerben B Keijzers, MB BS, MCLinEpi, FACEM, Emergency Department Physician,³ Associate Professor⁶

¹ ED Clinical Network, Gold Coast Hospital, and Griffith Health Institute, Queensland Health, Southport, QLD.

² Australian e-Health Research Centre, CSIRO ICT Centre, Brisbane, QLD.

³ Emergency Department, Gold Coast Hospital, Southport, QLD.

⁴ Australian Centre for Prehospital Research, Queensland Ambulance Service, and University of Queensland, Brisbane, QLD.

⁵ Griffith Health Institute, NHMRC Centre for Research Excellence in Nursing, and Gold Coast Health Service District, Southport, QLD.

⁶ Bond University, Southport, QLD.

Correspondence: julia_crilly@health.qld.gov.au

REFERENCES

- Queensland Health. The eHealth priorities, 2009. http://www.health.qld.gov.au/ehealth/ehealth_priorities.asp (accessed Feb 2010).
- Chan JK, Gomez SL, O'Malley CD, et al. Validity of cancer registry Medicaid status against enrollment files: implications for population-based studies of cancer outcomes. *Med Care* 2006; 44: 952-955.
- Kazanjian A. Understanding women's health through data development and data linkage: implications for research and policy. *CMAJ* 1998; 159: 342-345.
- Dean JM, Vernon DD, Cook L, et al. Probabilistic linkage of computerized ambulance and inpatient hospital discharge records: a potential tool for evaluation of emergency medical services. *Ann Emerg Med* 2001; 37: 616-626.
- Kelman CW, Bass AJ, Holman CD. Research use of linked health data — a best practice protocol. *Aust N Z J Public Health* 2002; 26: 251-255.
- Hall SE, Holman CD, Finn J, Semmens JB. Improving the evidence base for promoting quality and equity of surgical care using population-based linkage of administrative health records. *Int J Qual Health Care* 2005; 17: 415-420.
- Brook EL, Rosman DL, Holman CD. Public good through data linkage: measuring research outputs from the Western Australian data linkage system. *Aust N Z J Public Health* 2008; 32: 19-23.
- Spaite DW, Maio R, Garrison HG, et al. Emergency medical service outcomes project (EMSOP) II: Developing the foundation and conceptual models for out-of-hospital outcomes research. *Ann Emerg Med* 2001; 37: 657-663.
- Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records data in Western Australia: development of a health services research linked database. *Aust N Z J Public Health* 1999; 23: 453-459.
- Ives A, Saunders C, Bulsara M, Semmens J. Pregnancy after breast cancer: population based study. *BMJ* 2007; 334: 194.
- Sibthorpe B, Kliever E, Smith L. Record linkage in Australian epidemiological research: health

benefits, privacy safeguards and future potential. *Aust J Public Health* 1995; 19: 250-256.

- Boufous S, Finch C. Estimating the incidence of hospitalized injurious falls: impact of varying case definitions. *Inj Prev* 2005; 11: 334-336.
- 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.
- Cryer PC, Westrup S, Cook AC, et al. Investigation of bias after data linkage of hospital admissions data to police road traffic crash reports. *Inj Prev* 2001; 7: 234-241.
- Crilly J, Chaboyer W, Wallis M, et al. Predictive outcomes for older people who present to the emergency department. *Aust Emerg Nurs J* 2008; 11: 178-183.
- Karmel R, Anderson P, Gibson D, et al. Empirical aspects of record linkage across multiple data sets using statistical linkage keys: the experience of the PIAC cohort study. *BMC Health Serv Res* 2010; 10: 41-53.
- Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009; 9: 3-10.
- Hansen D, Pang C, Maeder A. HDI: integrating health data and tools. *Soft Comput* 2007; 11: 361-367.
- Bernstein SL, Aronsky D, Duseja R, et al. The effect of emergency department crowding on clinically orientated outcomes. *Acad Emerg Med* 2009; 16: 1-10.
- Sun BC, Mohanty SA, Weiss R, et al. Effect of hospital closures and hospital characteristics on emergency department ambulance diversion, Los Angeles County, 1998 to 2004. *Ann Emerg Med* 2006; 47: 309-316.
- Chu K. An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emerg Med* 1999; 11: 175-181.
- Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp* 2002: 305-309.
- Australian Institute of Health and Welfare. A national minimum data set for home and community care. Canberra: AIHW, 1999: 76. (AIHW Cat. No. AGE 13.) <http://www.aihw.gov.au/publications/index.cfm/title/4600> (accessed Dec 2010).
- Clark DE, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proc Annu Symp Comput Appl Med Care* 1995: 397-401.
- Australian Institute of Health and Welfare. Australian hospital statistics 2007-08. Canberra: AIHW, 2009. (Health services series No. 33; AIHW Cat. No. HSE 71.) <http://www.aihw.gov.au/publications/index.cfm/title/10776> (accessed Dec 2010).
- Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004; 363: 1147-1154.
- Moon G, Gould M, Jones K, et al. Epidemiology: an introduction. Philadelphia: Open University Press, 2000.

(Received 22 Mar 2010, accepted 1 Dec 2010) □