publichealth research&practice

September 2015; Vol. 25(4):e2541540
doi: http://dx.doi.org/10.17061/phrp2541540
www.phrp.com.au

Perspective

Routinely collected data as a strategic resource for research: priorities for methods and workforce

Louisa Jorm[a,b]

[a] Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia
[b] Corresponding author: l.jorm@unsw.edu.au

Article history

Publication date: September 2015
Citation: Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. Public Health Res Pract. 2015;25(4):e2541540. doi: http://dx.doi.org/10.17061/phrp2541540

Key points

- Research using routinely collected data can drive health system effectiveness and health improvement
- The policy environment increasingly supports the research use of routinely collected data
- Priorities for methods development include validation studies, and methods for analysing longitudinal data, exploring linkage error, and evaluation using 'natural experiments'
- Priorities for workforce development include building capabilities in computer science and research translation
- Large-scale, long-term partnership approaches involving government, industry and researchers offer the most promising way forward

Abstract

In the era of 'big data', research using routinely collected data offers greater potential than ever before to drive health system effectiveness and efficiency, and population health improvement. In Australia, the policy environment, and emerging frameworks and processes for data governance and access, increasingly support the use of routinely collected data for research. Capitalising on this strategic resource requires investment in both research methods and research workforce.

Priorities for methods development include validation studies, techniques for analysing complex longitudinal data, exploration of bias introduced through linkage error, and a robust toolkit to evaluate policies and programs using 'natural experiments'.

Priorities for workforce development include broadening the skills base of the existing research workforce, and the formation of new, larger, interdisciplinary research teams to incorporate capabilities in computer science, partnership research, research translation and the 'business' aspects of research.

Large-scale, long-term partnership approaches involving government, industry and researchers offer the most promising way to maximise returns on investment in research using routinely collected data.

Routinely collected data as a strategic resource

Large volumes of health-related data are collected routinely by governments, healthcare providers and insurers, as a byproduct of operating services.[1] Research and evaluation have traditionally been termed 'secondary' uses of these data, because the data are used for purposes other than those for which they were originally collected.

Other routinely collected data are generated to meet regulatory requirements (e.g. births, deaths, health practitioner registration), or specifically to monitor health conditions and health outcomes and to inform disease control and health improvement efforts (e.g. cancer notifications, communicable disease notifications, perinatal data). Statistical reporting,

research and evaluation have always been among the 'primary' uses of these data.

In the era of 'big data', the customary distinction between primary and secondary uses of routinely collected data is becoming irrelevant. Data are increasingly available in electronic form and can be linked over time, and across data sources, to create longitudinal records for individuals and multilevel data structures (e.g. patients within practices within geographic areas). Data systems such as clinical registries and electronic health records are being designed with both patient point-of-care and aggregate uses in mind.

The policy climate in Australia has increasingly supported greater research use of routinely collected data. The Council of Australian Governments released a National Government Information Sharing Strategy in 2009.[2] In 2010, the Australian Government endorsed a set of principles for data integration for research purposes, first among which is that agencies must treat data as a strategic resource and support their wider research use within and across policy portfolios.[3] Health is recognised as being a key beneficiary of increased use of routinely collected data for research, with a recent report from the Australian Productivity Commission stating that "there is a case for maximising the benefits that the community achieves from the data it has paid for. More extensive research and analysis of these data collections could deliver significant improvements in the efficiency and effectiveness of health care".[4] Recognising this, the National Health and Medical Research Council (NHMRC) has led the development of new national principles for the use of publicly funded data for research[5], which have been endorsed by key data custodians and are currently being finalised.

In this context, research using routinely collected data offers greater potential than ever before to drive health system effectiveness and efficiency, and population health improvement. Realising this potential will require capacity building in research methods and research workforce.

Priorities for research methods

Routinely collected data have particular advantages for research and evaluation, compared with bespoke modes of data collection[6], and they are also subject to specific limitations. These benefits and limitations, summarised in Table 1, indicate priorities for methodological research, development of new methods and new applications of existing methods.

Table 1. Benefits and limitations of using routinely collected health data for research and evaluation

Benefits	Limitations
Population reach: many routine data collections have whole-of-population coverage, and can be used to study rare outcomes (e.g. adverse events) and population subgroups (e.g. Indigenous Australians). They maximise power to identify even marginal shifts in practice as a result of new policies or programs.	**Event based**: many routine data collections contain records that are generated as the result of an event, such as a hospital episode or death. They provide no information about individuals who have not experienced such events, creating difficulties in defining denominators for calculation of rates or appropriate comparison groups.
Longitudinal: when linked internally or across datasets, routine data have a longitudinal structure that supports studies across the lifecourse, enabling long-term follow-up and allowing better causal inference.	**Uncertain validity**: valid recording of information (e.g. diagnoses) requires that the correct information is available for data entry (e.g. is present in the medical record) and that the correct value is entered. 'Rare' values in large datasets may be more likely to represent keystroke or coding errors than valid entries.
Avoids nonresponse, attrition and reporting bias: routine data have complete coverage, unlike surveys, which are subject to high and rising rates of socially and health-patterned nonresponse and attrition, as well as social desirability, reporting and recall biases.	**Limited data items**: routine data collections often contain only a parsimonious set of items relating to the administrative purpose of the collection, and as a result may include very limited information on key confounders and risk factors, such as socioeconomic status, smoking status, bodyweight or height.
Cost-effective: the use of routine data for research and evaluation increases return on investment for the public resources expended in collecting them, and studies of whole populations over many decades can be undertaken time-efficiently and cost-efficiently compared with prospective data collection.	**Linkage error**: bias may be introduced through errors in linkage, which are likely to be nonrandom (e.g. difficulties in matching names may be more prevalent in people from some cultural backgrounds).
Real world: routine data often present the only way to evaluate the outcomes of care in population groups for which there is no evidence from clinical trials, or to evaluate the impacts of policies or services that have been rolled out in a nonrandomised manner.	**Lack of metadata**: detailed metadata and documentation to support the research use of the data may not be readily, or publicly, available.

An increasing number of validation studies have investigated the accuracy of recording of demographics, diagnoses, comorbidities and procedures in routinely collected Australian hospital[7–10] and perinatal[11] data. There is an ongoing need for such studies. Data quality is fluid and will change with coding standards and practices, the evolution of information systems, and sometimes in response to policy changes (e.g. 'upcoding' in response to casemix-based funding).[12]

Robust methods for analysis of longitudinal data are needed for use with linked routinely collected data and emerging routine data sources that are inherently longitudinal and person-centred, such as electronic medical records. Priorities include techniques for dealing with missing values[13], methods to explore the impact of measurement error and unmeasured confounders[14], joint modelling of survival and longitudinal data[15], methods for modelling recurrent events[16] and methods for real-time data analysis to support clinical prediction.[17] Additionally, new ways to visualise complex health data for large numbers of individuals[18] hold promise for both exploration of data and communication of research findings to policy makers and the public.

The use of probabilistic ('fuzzy') matching to link routine data in Australia (many other countries mainly rely on unique health identifiers for data linkage) creates an imperative to explore the potential bias that may be introduced through linkage error.[19] This may be especially important in large routine data studies, which are better protected from sampling (random) error because of their size, but are not protected from sources of systematic error. Data linkage errors are likely to be nonrandom – for example, difficulties in matching names may be more prevalent in people from some cultural backgrounds, and healthier people may be more likely to migrate and therefore be lost to follow-up through linkage.

Perhaps most pressing of these methodological priorities is the need to develop a robust toolkit for evaluation of policies and programs using 'natural experiments'.[6] Natural experiments use changes in programs and services that are not randomised, but they are the next best thing to a randomised controlled trial (RCT). They can be an extremely valuable and practical (and often the only) means of evaluating changes in policies or services. The key to making causal inference about the policy or service change is finding a reasonable comparison group that has not been exposed to the change but is sufficiently similar to the exposed population, analogous to the 'untreated' arm of an RCT. Methods such as propensity score matching, inverse probability of treatment weighting and interrupted time-series analysis are increasingly being applied to evaluate policies and programs[6], but a common problem is a lack of robust information about the characteristics of these 'interventions', and of how, when and where they were implemented.

Priorities for research workforce development

Despite the recent investment in infrastructure and development of policy frameworks to support research using routinely collected data, there has been little matching investment in expanding Australia's human capacity to do this research.

A notable exception is the NSW Ministry of Health's Biostatistics Training Program, which has been operating since 2000. This is a 3-year training program in which trainees rotate through a series of work placements and study part time, through distance learning, for a Master of Biostatistics degree from the Biostatistics Collaboration of Australia. Almost 50 trainees have completed the program, and demand for graduates is high. Additionally, the NHMRC has designated biostatistics and bioinformatics as priority areas for its people support schemes since at least 2012, and encourages applications from these disciplines.

However, the workforce capabilities required for health research using routinely collected data extend well beyond biostatistics and bioinformatics (summarised in Table 2). These imply both a broadening of the skills base of the existing research workforce and the formation of new, larger, interdisciplinary research teams.

Table 2. Workforce capabilities for research using routinely collected health data

Skills in	Knowledge of
Applying biostatistical methods	Bioinformatics
Conference presentations	Biostatistics
Data management	Communications
Data manipulation	Computer programming
Data security	Computer science
Database design	Data governance
Grantsmanship	Data provenance and interpretation
Literature review and synthesis	Data security and privacy
Managing contracts	Data standards
Project management	Data structures
Research design	Epidemiology
Specifying research questions	Ethics
Visualisation design	Health and clinical domains
Working with databases	Health system structure and operation
Working with policy partners	Machine learning
Working with the media	Meta-analysis
Writing blogs and commentaries	Metadata standards
Writing ethics applications	Research methodology
Writing for policy	Research translation
Writing grant applications	Social media
Writing scientific papers	Unstructured data (e.g. images, text)
Writing tenders	Visualisation

Despite concern (expressed by statisticians) that statistics is 'losing ground' to computer science and that this implies a less rigorous approach[20], few would deny that efficiencies derived from computer science are underused in health research using routinely collected data. This research still operates largely under a project-based, 'cottage industry' approach, with limited sharing of data and methods, and fails to rapidly take advantage of new technologies to increase computational efficiency, safeguard data security, and overcome legal and jurisdictional barriers to data pooling.

On the other hand, appropriate application of computer science approaches such as machine learning and data mining in health research requires that content domain expertise is used. Researchers using routinely collected data need sophisticated skills in working with clinicians, health services and policy makers to identify the key research questions, answer these in the right way, and facilitate the uptake of findings into policy and practice. They also need a detailed understanding of the strengths and weaknesses of routinely collected data, including issues of provenance, quality, and changes in data standards and coding practices, to avoid making novice mistakes in interpreting their findings.

Other priority capability areas such as 'grantsmanship' and project and contract management relate to the 'hand-to-mouth' nature of health and medical research in Australia, with intense competition for limited grant income, and short-term employment tenure as a norm. Researchers now need to be adept at the 'business' of research. Unfortunately, even contract research commissioned by government agencies is usually project based rather than programmatic in nature. Again, this is very inefficient, with funds and effort expended on the lengthy start-up approval and set-up phases of new research involving routinely collected data, rather than on maximising the knowledge outputs.

Conclusions

Making best use of routinely collected data to drive improvement in health services and health is a national priority for Australia.[4] The policy environment, and emerging frameworks and processes for data governance and access, increasingly support the use of routinely collected data for research. Capitalising on this strategic resource requires investment in both research methods and research workforce. Large-scale, long-term partnership approaches involving government, industry and researchers offer the most promising way to maximise the returns on such investment. Reliance on competitive research grant funding mechanisms that are increasingly driven by citation metrics, which do not favour research that focuses on local health systems, is very unlikely to achieve this.

Competing interests

None declared

Author contributions

LJ is sole author.

References

1. Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using English administrative data (hospital episode statistics) to assess health-care outcomes – systematic review and recommendations for reporting. Eur J Public Health. 2013;23(1):86–92.
2. Australian Government Information Management Office. National government information sharing strategy: unlocking government information assets to benefit the broader community. Canberra: Australian Governments; 2009 [cited 2015 Aug 12]. Available from: www.finance.gov.au/files/2012/04/ngiss.pdf
3. National Statistical Service. High-level principles for data integration involving Commonwealth data for statistical and research purposes. Canberra: Australian Government; 2010 [cited 2015 Aug 12]. Available from: www.nss.gov.au/nss/home.NSF/pages/High+Level+Principles+for+Data+Integration+-+Content?OpenDocument
4. Productivity Commission. Public and private hospitals: multivariate analysis. Supplement to research report. Canberra: Australian Government; 2010 [cited 2015 Aug12]. Available from: www.pc.gov.au/inquiries/completed/hospitals/supplement
5. National Health and Medical Research Council. Targeted consultation on the draft principles for accessing and using publicly funded data for health research. Canberra: Australian Government; 2014 [cited 2015 Aug 12]. Available from: consultations.nhmrc.gov.au/public_consultations/funded-data
6. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. J Epidemiol Community Health. 2012;66(12):1182–6.
7. Tran D, Jorm L, Lujic S, Bambrick H, Johnson M. Country of birth recording in Australian hospital morbidity data: accuracy and predictors. Aust N Z J Public Health. 2012;36(4):310–6.
8. Randall DA, Lujic S, Leyland AH, Jorm LR. Statistical methods to enhance reporting of Aboriginal Australians in routine hospital records using data linkage affect estimates of health disparities. Aust N Z J Public Health. 2013;37(5):442–9.

9. Lujic S, Watson DE, Randall DA, Simpson J, Jorm LR. Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. BMJ Open. 2014;4(9):e005768.
10. Tran DT, Roberts CL, Havard A, Jorm LR. Linking birth records to hospital admission records enhances the identification of women who smoke during pregnancy. Aust N Z J Public Health. 2014;38(3):258–64.
11. Lain SJ, Hadfield RM, Raynes-Greenow CH, Ford JB, Mealing NM, Algert CS, Roberts CL. Quality of data in perinatal population health databases: a systematic review. Med Care. 2012;50:e7–20.
12. Steinbusch PJ, Oostenbrink JB, Zuurbier JJ, et al. The risk of upcoding in casemix systems: a comparative study. Health Policy. 2007;81(2–3):289–99.
13. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodol. 2012;12:96.
14. Groenwold RHH, Nelson DB, Nichol KL, Hoe AW, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. Int J Epidemiol. 2010;39(1):107–17.
15. Wu L, Liu W,Yi GY, Huang Y. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. J Probability and Statistics. 2012:640153.
16. Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324–33.
17. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. Journal of Big Data. 2014;(1):2.
18. Monroe M, Lan R, Plaisant C, Shneiderman B. Temporal event sequence simplification. In IEEE Trans Visualization and Computer Graphics. 2013;19(12):2227–36.
19. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Med Res Methodol. 2014;14:36.
20. Matloff N. Statistics is losing ground to computer science. London: Royal Statistical Society; 2014 [cited 2015 Aug 12]. Available from: www.statslife.org.uk/opinion/1887-statistics-is-losing-ground-to-computer-science