publichealth research&practice

July 2016; Vol. 26(3):e2631636
doi: http://dx.doi.org/10.17061/phrp2631636
www.phrp.com.au

In practice

Data linkage in an established longitudinal cohort: the Western Australian Pregnancy Cohort (Raine) Study

Jenny A Mountain[a,e], Anett Nyaradi[a,b], Wendy H Oddy[c], Rebecca A Glauert[b], Nick H de Klerk[b], Leon M Straker[d] and Fiona J Stanley[b]

[a] School of Population Health, The University of Western Australia, Perth
[b] Telethon Kids Institute, The University of Western Australia, Perth
[c] Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia
[d] School of Physiotherapy and Exercise Science, Curtin University, Perth, Western Australia
[e] Corresponding author jenny.mountain@uwa.edu.au

Article history

Publication date: July 2016
Citation: Mountain JA, Nyaradi A, Oddy WH, Glauert RA, de Klerk NH, Straker LM, Stanley FJ. Data linkage in an established longitudinal cohort: the Western Australian Pregnancy Cohort (Raine) Study. Public Health Res Pract. 2016;26(3):e2631636. doi: http://dx.doi.org/10.17061/phrp2631636

Key points

- Data linkage merges datasets while retaining the privacy of individuals
- The Raine Study is an established longitudinal cohort study with detailed information collected about more than 2000 participants for 26 years
- Educational outcomes data held by government were linked to the Raine Study dataset
- Analysis of these data found important associations between eating a healthy diet and achieving higher educational outcomes
- The data linkage process is lengthy, but it establishes required approvals and linkage keys, which improves research opportunities

Abstract

The Western Australian Data Linkage System is one of a few comprehensive, population-based data linkage systems worldwide, creating links between information from different sources relating to the same individual, family, place or event, while maintaining privacy. The Raine Study is an established cohort study with more than 2000 currently active participants.

Individual consent was obtained from participants for information in publicly held databases to be linked to their study data. A waiver of consent was granted where it was impracticable to obtain consent. Approvals to link the datasets were obtained from relevant ethics committees and data custodians. The Raine Study dataset was subsequently linked to academic testing data collected by the Western Australian Department of Education.

Examination of diet and academic performance showed that children who were predominantly breastfed for at least 6 months scored higher academically at age 10 than children who were breastfed for less than 6 months. A further study found that better diet quality at ages 1, 2 and 3 years was associated with higher academic scores at ages 10 and 12 years. Examination of nutritional intake at 14 years of age found that a better dietary pattern was associated with higher academic performance. The detailed longitudinal data collected in the Raine Study allowed for adjustment for multiple covariates and confounders.

Data linkage reduces the burden on cohort participants by providing additional information without the need to contact participants. It can give information on participants who have been lost to follow-up; provide or complement missing data; give the opportunity for validation studies comparing recall of participants with administrative records; increase the population sample of studies by adding control participants from the general

population; and allow for the adjustment of multiple covariates and confounders. The Raine Study dataset is extensive and detailed, and can be further improved by linking to other external data sources. By linking educational outcomes to the Raine Study database, it was shown across three different age groups that a healthy diet was consistently associated with higher academic performance.

Introduction

The Western Australian Pregnancy Cohort (Raine) Study was established with the enrolment of 2900 pregnant women at 18 weeks gestation attending a public antenatal clinic and nearby private clinics in Perth, Australia, between May 1989 and November 1991.[1] The study's purpose was to investigate whether intensive use of ultrasound imaging and Doppler flow studies would improve pregnancy outcomes, and to develop a long-term cohort to examine the role of early life events on later health.[2] Extensive data were collected during pregnancy, and 2868 offspring were assessed at birth and followed up at 1, 2, 3, 5, 8, 10, 14, 17, 18, 20 and 22 years of age.[3] Questionnaire data, physical measurements and biological samples were collected during pregnancy and infancy, resulting in a database containing a broad range of prospective demographic data, detailed phenotype data, and measures of the antenatal and postnatal environment. Data collection was expanded during childhood, adolescence and young adulthood to cover a lifecourse framework, including phenotype, environmental, behavioural, occupational and genetic information. Information routinely collected in administrative databases (such as hospital admissions, medication use and education outcomes) has the potential to be linked to the existing Raine Study data. The purpose of this article is to outline the potential benefit to longitudinal cohorts of linkage to administrative datasets, using the first Raine Study data linkage involving diet (collected in the Raine Study at follow-up) and educational outcome (collected by the local education authority) as an example.

Begun in 1995 to link local health datasets, the current Western Australian Data Linkage System (WADLS) is one of the few comprehensive, population-based data linkage systems worldwide.[3] Data linkage is a technique for creating links between information from different sources that relate to the same individual, family, place or event.[3] Data custodians are the organisations or agencies (or their representatives) responsible for the collection and use of datasets, and have access to identifying demographic and phenotypic information. Data custodians are responsible for protecting the privacy of individuals, according to both legislation and public interest in the right to privacy of personal information.[4] The Raine Study Executive Committee is the custodian for Raine Study data.

Data are separated into two distinct types: identifying information (e.g. name, address, date of birth), and clinical or service information (e.g. hospital admissions or educational records). Data custodians provide identifying information to the WADLS for linkage using a probabilistic matching approach, and clerical review when necessary. Linkage keys, unique to each individual, are generated across different data collections.[5] The linkage keys are encrypted and replace all identifying information. The encrypted identifiers are returned to the data custodians, who provide service information with the encrypted keys, but no identifying information, to the researcher.

The WADLS acts as an intermediary between data custodians and researchers. Western Australian Department of Health (WADoH) Human Research Ethics Committee (HREC) approval is required for creating new linkages and using the linkage infrastructure. Separate approvals and agreements are obtained from all agencies and data custodians before linkage can be conducted. Researchers receiving linked data agree, in writing, to a strict set of data security conditions.[3]

Data linkage is generally used for large, population-based projects, facilitating merging of large sources of information on individuals. For example, Silva et al. examined 12 831 individuals aged 10–21 years with a diagnosis of attention deficit hyperactivity disorder (ADHD) who were matched by age, sex and socio-economic status with a further 29 722 individuals with no ADHD diagnosis.[6] Data linkage also enables linking of trial data to other datasets. For example, Kelty et al. linked self-reported adverse events from a clinical trial with hospital mortality data and emergency data collections.[7] This also allowed for the validation of self-reported events against hospital admissions data. Longitudinal cohort data can also be linked to other datasets – for example, Knuiman et al. linked the Busselton Cohort Study to death records to ascertain mortality rates over a 26-year period.[8]

The potential to link Raine Study information with data from more than 30 collections – including local and national health and welfare datasets, genealogical links and spatial references[3] – can significantly complement and improve the research potential of the Raine Study. As a first step and proof of principle, Raine Study data have been successfully linked with administrative records managed by the WADLS. The success has been in terms of both logistically linking the data and providing useful findings. This linkage enabled researchers to examine the associations between Raine Study diet data and educational outcome data collected by the Western Australian Department of Education.[9] Here, we outline the steps necessary to conduct the linkage, its value and the challenges faced.

Methods

For child participation in the Raine Study, written informed consent was obtained from parents or guardians at

pregnancy, birth, and at 1, 2, 3, 5, 8, 10, 14 and 17 years of age. Assent was obtained from the participants at age 14 and 17. In 2008, when the cohort participants started to turn 18 years old, approval was granted by the University of Western Australia HREC to contact all Raine Study 18-year-old participants (including participants whose parents had previously withdrawn), and acquire consent for the Raine Study to:

- Contact them for future assessments
- Use their previously collected data, including DNA and stored biological samples
- Obtain additional information through data linkage with administrative records.

Where the Raine Study failed to contact the participant, or the participant did not respond, a waiver of consent (sought to protect, maintain, maximise and improve value of the data previously collected on the cohort) was granted by the University of Western Australia Human Research Ethics Committee.

Information and consent forms were sent to more than 2500 cohort members who were not lost to follow-up, withdrawn or deceased (Figure 1). Over 12 months, 1127 participants returned a signed consent form. Of these, five declined permission for data linkage and were not included when creating linkage keys, and 949 active participants did not return a signed form. Under the waiver of consent, those lost to follow-up and nonresponders, including withdrawn participants, were included for data linkage.

In 2012, the Raine Study made a successful application to the WADoH HREC (#2012/70) and obtained approval for linkage to the WADLS.

Specific project applications to link the Raine Study longitudinal diet data with educational results were made to the Raine Study Executive Committee, the WADLS (ref #2013/75) and the WA Developmental Pathways Project (WADPP). The WADPP includes nonhealth custodians, which enables linkage of a number of nonhealth datasets to the WADLS, including education.[10] The participant consent and approvals processes, and data manipulation and labelling took 1 year to complete.

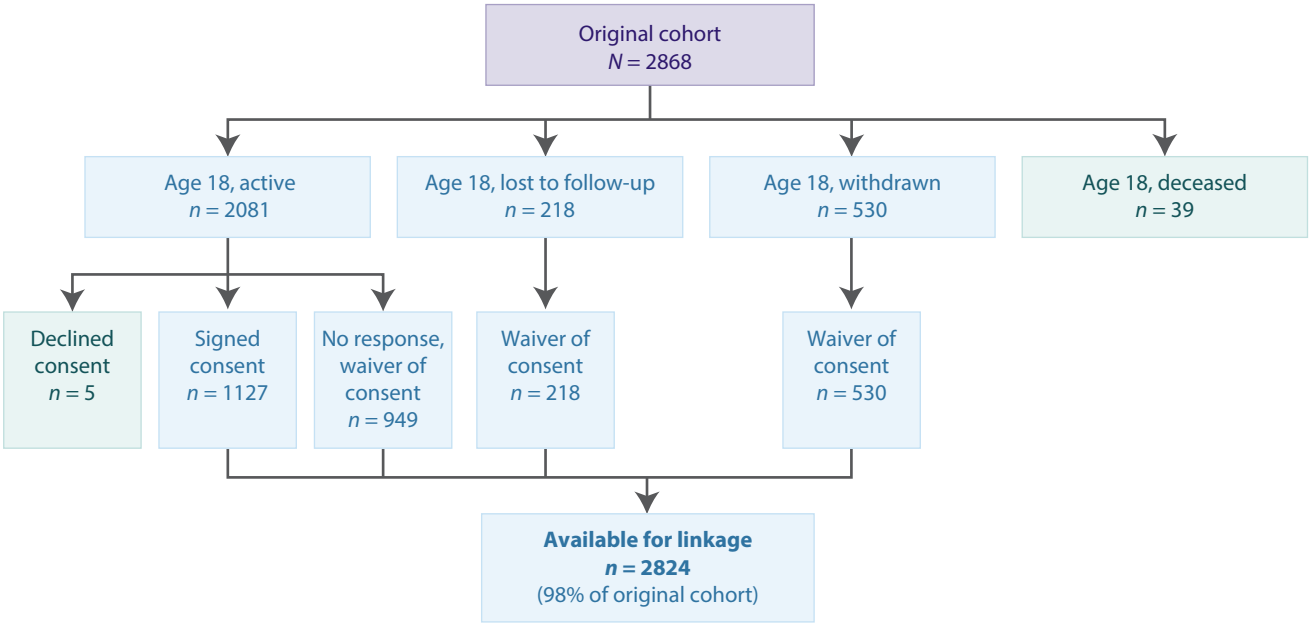
Results

Raine Study data, with information on diet in infancy, childhood and adolescence, were linked to Western Australian Literacy and Numeracy Assessment (WALNA) records for participants in Years 5 (age 10), 7 (age 12) and 9 (age 14).

The first study examined the relationship between duration of breastfeeding (available from the cohort at birth, 1, 2 and 3 years, based on the maternal report) and WALNA scores in Year 5 (age 10 years, $n = 1038$).[11] Children who had been predominantly breastfed for at least 6 months attained better educational outcome scores (based on WALNA results at mean age 10.4 years) in mathematics (β 15.79; 95% confidence interval [CI] 1.04, 30.55; $p = 0.036$) and reading (β 18.28; 95% CI 3.92, 32.64; $p = 0.021$) in the multivariable model than those breastfed for less than 6 months, after adjusting for marital status, maternal age and education, early reading with child and household income. The effect was gender-specific, with only boys who had been breastfed for at least 6 months attaining significantly better scores in mathematics (β 34.48; 95% CI 13.66, 55.27; $p = 0.001$), reading (β 24.72; 95% CI 4.87, 44.58; $p = 0.015$), writing (β 37.41; 95% CI 9.43, 65.40; $p = 0.009$) and spelling (β 28.23; 95% CI 6.38, 50.07; $p = 0.011$).

The second study examined the relationship between early diet and WALNA scores in Year 5 (age 10, $n = 2247$)

Figure 1. Flowchart of Raine Study participant consent for data linkage at age 18



and Year 7 (age 12, n = 2287).[12] A modified 24-hour dietary recall was completed by the parent or guardian at ages 1, 2 and 3. An overall dietary score was developed, with higher scores representing greater intakes of fruits, vegetables, wholegrains and protein sources (excluding processed and red meats), and lower intake of sweetened drinks and snack foods.[12,13] The extensive Raine Study phenotype dataset allowed for adjustment of the effects of gender; maternal age, race and education; family income; the presence of the biological father; breastfeeding duration; and parental language stimulation (reading to the child) from birth to age 3. A better-quality diet at 1 year old was associated with significantly higher WALNA scores in Year 5 (mathematics [β 0.47; 95% CI 0.09, 0.84; p = 0.015], reading [β 0.63; 95% CI 0.24, 1.02; p = 0.002], writing [β 0.94; 95% CI 0.37, 1.50; p = 0.001] and spelling [β 0.89; 95% CI 0.31, 1.47; p = 0.003]) and Year 7 (mathematics [β 0.71; 95% CI 0.25, 1.16; p = 0.002], reading [β 0.75; 95% CI 0.39, 1.10; p < 0.001] and spelling (β 0.90; 95% CI 0.35, 1.45; p = 0.001]).

The third study examined the association between dietary patterns in adolescence and WALNA results in Year 9 (age 14, n = 779 mathematics, n = 741 reading and n = 470 writing).[7] At age 14, dietary data were collected using a parent-completed food frequency questionnaire.[14] Dietary patterns were derived by factor analysis from the major food groups.[15] Two major dietary patterns were identified: 'healthy' (high intake of fruits, vegetables, wholegrains, legumes and fish) and 'nonhealthy' (high intake of takeaway foods, red and processed meat, soft drinks, fried and refined food). Adjustments were made for sociodemographic and family characteristics, and it was found that higher scores for the nonhealthy dietary pattern were significantly associated with poorer WALNA scores for mathematics (β −13.14; 95% CI −24.57, −1.76; p = 0.024) and reading (β −19.16; 95% CI −29.85, −8.47; $p \leq 0.001$), with a similar trend found for writing (β −17.28; 95% CI −35.74, 1.18; p = 0.066).

Discussion

Linking the Raine Study data with other datasets was a laborious and lengthy – but ultimately worthwhile – process. Findings from linking the education outcomes to the Raine Study diet data provided clear evidence of a link between healthy dietary intake and academic achievement across three lifecourse age periods. The initial efforts we made to obtain HREC approval, individual consent from cohort participants, a waiver of consent for linkage, ethical approval and data custodian permission to establish the linkage keys will pave the way for future data linkage project applications.

Data linkage with other administrative data serves to reduce the burden on participants. Data collection methods within the Raine Study predominantly require participant involvement, including self-report questionnaires, diaries, interviews, physical examination, clinical testing and biological sampling. The current linkage allowed accurate educational outcomes to be ascertained without burdening the participants, and without introducing error from participant recall mistakes and missing data from participant failure to respond to or recall the data. Where participants were lost to follow-up, linkage with existing datasets can provide missing information on outcomes such as hospital admissions. Data not collected in previous follow-ups can be sourced. If the measurement of an exposure or outcome is available from more than one source, data linkage can help achieve optimal accuracy of data by comparing and validating the measures used.[16] Linkage to total population data also allows comparison of the cohort with patterns in the population to assess representativeness.

Linking the Raine Study data with other datasets, including those available through the WADLS (from collections concerning birth notifications, mental health, child protection, disability, hospital admissions, education, case registries [diabetes, cancer, autism], deaths, medication use and primary care records), can add information to an existing rich database and expand the use of the Raine Study data for cross-disciplinary research projects. The high-quality findings provided by longitudinal studies with linked data, as outlined in this paper, are important sources of evidence for health and social policy guidelines.

Acknowledgements

We acknowledge the Raine Study participants and their families. The following institutions are acknowledged for core funding: the University of Western Australia (UWA); Curtin University; the Telethon Kids Institute; the Raine Medical Research Foundation; the UWA Faculty of Medicine, Dentistry and Health Sciences; the Women's and Infant's Research Foundation; and Edith Cowan University. We acknowledge the WADLS for linkage support, the WADPP for enabling linking to a number of nonhealth datasets (supported by Australian Research Council Linkage Project 100200507) and the Western Australian Department of Education for providing the educational data.

Competing interests

None declared

Author contributions

JM designed, drafted and edited the manuscript, and analysed the data; AN and WO reviewed and edited the manuscript; RG designed, drafted and edited the manuscript; NK provided analytical advice, and reviewed and edited the manuscript; LS provided analytical advice, and drafted, reviewed and edited the manuscript; and FS reviewed and edited the manuscript, and contributed to its design.

References

1. Newnham JP, Evans SF, Michael CA, Stanley FJ, Landau LI. Effects of frequent ultrasound during pregnancy: a randomised controlled trial. Lancet. 1993;342(8876):887–91.
2. McKnight CM, Newnham JP, Stanley FJ, Mountain JA, Landau LI, Beilin LJ, et al. Birth of a cohort – the first 20 years of the Raine study. Med J Aust. 2012;197(11):608–10.
3. Holman CD, Bass JA, Rosman DL, Smith MB, Semmens JB, Glasson EJ, et al. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. Aust Health Rev. 2008;32(4):766–77.
4. Kelman CW, Bass AJ, Holman CD. Research use of linked health data – a best practice protocol. Aust N Z J Public Health. 2002;26(3):251–5.
5. Holman CD, Bass AJ, Rouse IL, Hobbs MS. Population-based linkage of health records in Western Australia: development of a health services research linked database. Aust N Z J Public Health. 1999;23(5):453–9.
6. Silva D, Colvin L, Glauert R, Bower C. Contact with the juvenile justice system in children treated with stimulant medication for attention deficit hyperactivity disorder: a population study. Lancet Psychiatry. 2014;1(4):278–85.
7. Kelty E, Ngo H, Hulse G. Assessing the usefulness of health data linkage in obtaining adverse event data in a randomised controlled trial of oral and implant naltrexone in the treatment of heroin dependence. Clin Trials. 2013;10(1):170–80.
8. Knuiman MW, James AL, Divitini ML, Ryan G, Bartholomew HC, Musk AW. Lung function, respiratory symptoms, and mortality: results from the Busselton Health Study. Ann epidemiol. 1999;9(5):297–306.
9. Nyaradi A, Li J, Hickling S, Foster JK, Jacques A, Ambrosini GL, Oddy WH. A Western dietary pattern is associated with poor academic performance in Australian adolescents. Nutrients. 2015;7(4):2961–82.
10. Stanley F, Glauert R, McKenzie A, O'Donnell M. Can joined-up data lead to joined-up thinking? The Western Australian developmental pathways project. Healthc Policy. 2011;6(Spec Issue):63–73.
11. Oddy WH, Li J, Whitehouse AJ, Zubrick SR, Malacova E. Breastfeeding duration and academic achievement at 10 years. Pediatrics. 2011;127(1):e137–45.
12. Nyaradi A, Li J, Foster JK, Hickling S, Jacques A, O'Sullivan TA, Oddy WH. Good quality diet in the early years may have a positive effect on academic achievement. Acta Paediatr. 2016;105(5):e209–18.
13. Nyaradi A, Li J, Hickling S, Whitehouse AJ, Foster JK, Oddy WH. Diet in the early years of life influences cognitive outcomes at 10 years: a prospective cohort study. Acta Paediatr. 2013;102(12):1165–73.
14. Baghurst KI, Record SJ. A computerised dietary analysis system for use with diet diaries or food frequency questionnaire. Community Health Stud. 1984;8(1):11–8.
15. Ambrosini GL, Oddy WH, Robinson M, O'Sullivan TA, Hands BP, de Klerk NH, et al. Adolescent dietary patterns are associated with lifestyle and family psycho-social factors. Public Health Nutr. 2009;12(10):1807–15.
16. Golding J, Jones R. Sources of data for a longitudinal birth cohort. Paediatr Perinat Epidemiol. 2009;23 Suppl 1:51–62.

Copyright:

© 2016 Mountain et al. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Licence, which allows others to redistribute, adapt and share this work non-commercially provided they attribute the work and any adapted version of it is distributed under the same Creative Commons licence terms. See: www.creativecommons.org/licenses/by-nc-sa/4.0/