Potential for advances in data linkage and data science to support injury prevention research

Ronan A Lyons [1,2,3] Belinda J Gabbe,[1,2] Kirsten Vallmuur [4,5]

[1]Population Data Science, Swansea University, Swansea, UK
[2]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
[3]Administrative Data Research Wales, Swansea University Medical School, Swansea University, Swansea, UK
[4]Australian Centre for Health Services Innovation (AusHSI), Queensland University of Technology (QUT), Brisbane, Queensland, Australia
[5]Jamieson Trauma Institute, Royal Brisbane & Women's Hospital (RBWH), Brisbane, Queensland, Australia

Correspondence to
Professor Ronan A Lyons; r.a.lyons@swansea.ac.uk

Received 15 May 2024
Accepted 14 September 2024
Published Online First 3 October 2024

To cite: Lyons RA, Gabbe BJ, Vallmuur K. *Inj Prev* 2024;**30**:442–445.

ABSTRACT

The recent COVID-19 pandemic stimulated unprecedented linkage of datasets worldwide, and while injury is endemic rather than pandemic, there is much to be learned by the injury prevention community from the data science approaches taken to respond to the pandemic to support research into the primary, secondary and tertiary prevention of injuries. The use of routinely collected data to produce real-world evidence, as an alternative to clinical trials, has been gaining in popularity as the availability and quality of digital health platforms grow and the linkage landscape, and the analytics required to make best use of linked and unstructured data, is rapidly evolving. Capitalising on existing data sources, innovative linkage and advanced analytic approaches provides the opportunity to undertake novel injury prevention research and generate new knowledge, while avoiding data waste and additional burden to participants. We provide a tangible, but not exhaustive, list of examples showing the breadth and value of data linkage, along with the emerging capabilities of natural language processing techniques to enhance injury research. To optimise data science approaches to injury prevention, injury researchers in this area need to share methods, code, models and tools to improve consistence and efficiencies in this field. Increased collaboration between injury prevention researchers and data scientists working on population data linkage systems has much to offer this field of research.

INTRODUCTION

The recent COVID-19 pandemic stimulated unprecedented linkage of datasets in many jurisdictions to better understand the patterns of transmission of the SARS-CoV-2 virus, vulnerability to disease and the effectiveness of counter-measures.[1] While injury is endemic rather than pandemic, there is much to be learnt by the injury prevention community from the approaches taken to respond to the pandemic. Using examples, we highlight how various aspects of data science techniques, such as data linkage, and natural language processing (NLP) can support research into the primary, secondary and tertiary prevention of injuries. This can, in turn, provide a more comprehensive understanding of risk factors and counter-measures to inform programme and policy development.

Data linkage is a technique in which data from an individual or an entity (such as a household or organisation) are linked together. Pieces of information about individuals are collected by many different organisations for health, work, education, taxation and many other purposes and exist in large numbers of unlinked data siloes. Bringing such data together poses many challenges, not least privacy protection. However, this is possible using a common unique identity for the individual or entity held in the different databases or through probabilistic matching based on a variety of attributes such as name, sex, date of birth and address. The Administrative Data Research UK website provides a good example of how such identity matching is achieved.[2] In almost all jurisdictions, this is legally possible and is increasingly successfully conducted through the creation and utilisation of trusted research environments (TREs), which hold deidentified data from many sources on the general population. More details on the technical aspects behind these developments and their widespread use can be found in a variety of online sources such as the International Population Data Linkage Network (https://ipdln.org/) and the video produced by the Australian Population Health Research Network.[3]

Why use of data linkage in injury research?

Given the well-known lack of funding for injury prevention research[4] and the expense of conducting large scale and long-term randomised trials in the community, there is a dearth of trial-based evidence for many promising interventions. The use of routinely collected data to produce real-world evidence, as an alternative to clinical trials, has been gaining in popularity as the availability and quality of digital health platforms grow.[5]

Arguably, given the social drivers of injury risk, such as the underlying differential exposures to hazards and the ability to modify and respond to these as experienced by people from different socioeconomic, racial, ethnic, geographic and income groups,[6] many interventions aimed at non-injury factors could prevent injuries or reduce their severity, but these data are often lacking or insufficiently characterised when single data sources are used. Embedding trials, cohorts and evaluations of natural experiments in population data linkage systems with a wide array of health and other data provides opportunities to conduct evaluations with respect to injury outcomes and a wide range of their consequences across individual, family and societal domains as highlighted in the Injury List of All Deficits (LOAD) framework,[7] even when these were not originally planned. In addition, extensive data are captured in text-based form on the antecedents of injury in paramedic case descriptions, emergency department (ED) presenting complaint and triage fields, hospital admission records and discharge

summaries, and in other routinely completed specialist forms and notes along a person's treatment journey. Radiology reports and other clinical records contain rich detail on the nature and severity of injury, much of which is not incorporated in routine coding systems such as International Classification of Diseases 10th revision. Similarly, police crash records, insurance records and ED text narratives of the injury event contain valuable information about the circumstances of the event, built environment and road infrastructure and safety.[8] Very little of this is coded and available in electronic sources accessible by researchers. NLP has the ability to improve the quality and depth of data capture, which when linked to other datasets, can substantially improve the utility and completeness of data. NLP is based out of the fields of artificial intelligence and linguistics and allows computers to interpret words and phrases written by people and includes named entity recognition and text summarisation which could help autoencode diagnoses and mechanisms of injury and underlying aetiology, as we detail below.[9] Recently, there has been considerable interest in the development of a component of NLP known as large language models (LLMs), which are computer models capable of generating text and predicting answers and are widely used in internet search engines and predictive text.

The linkage landscape, and the analytics required to make best use of linked and unstructured data, is rapidly evolving. What might not have been possible previously can become feasible quite quickly. Capitalising on existing data sources, innovative linkage and advanced analytic approaches provides the opportunity to undertake novel injury prevention research, and generate new knowledge, while avoiding data waste and additional burden to participants.

Key uses of data linkage in injury prevention research

Population-based data linkage provides the potential to evaluate:

1. Embedded individual or cluster randomised trials which link to injury outcomes.
2. Cohorts and surveys which links to injury outcomes.
3. Evaluations of natural experiments with links to injury outcomes.

Embedded randomised trials

Notably, injury prevention or injury treatment trials with long-term outcomes monitored by linkage to routine data are largely absent. This reflects a wider trend. In the UK, where record linkage is well established, only a minority (<3%) of trials are linked to routine data.[10] Nevertheless, studies have shown the capability of data linkage to establish important end points in trials. For example, the West of Scotland Coronary Prevention Study highlighted the benefits of long-term data linkage.[11] In this randomised, placebo-controlled, primary prevention trial of pravastatin (a medication designed to lower cholesterol in the blood), capturing long-term mortality through data linkage dramatically changed the cost benefit analysis of statin prescribing for heart disease prevention. The 15-year linkage was not planned in the original 5-year study and cost a mere £15 000 and is arguably the best example of the cost-effectiveness of record linkage research.[11]

Cohort studies

The use of population-based cohort studies using record linkage to evaluate injury risk and outcomes is more prevalent. The Millennium Cohort Study (MCS) in the UK identified an association between physical activity and injury risk; boys, but not girls, who were overall more physically active experienced higher rates of injuries resulting in ED attendances and hospital admissions.[12 13] The Avon Longitudinal Study of Parents and Children birth cohort demonstrated that much of area-level variations in childhood injuries was due to variations in maternal and child health risk factors in individuals clustered into neighbourhoods.[14 15] Several population-based cohort studies have identified the detrimental impact of childhood injury admission on educational attainment.[16–18] In the USA, the MCS was set up to evaluate the impact of military experience on service members and veteran health using a recruited cohort 260 228 military personnel across 5 panels between 2001 and 2021, with repeated surveys and data linkage to administrative and medical data sources, which has enabled many studies including the relationship between deployment injuries and mental and physical quality of life.[19]

A number of population cohort studies using linked data have focused on priority and at-risk populations, which have been historically challenging populations to study. Using data linkage in Ontario, Canada, O'Neill *et al* were able to explore mental health and assault care treatment in the year prior to death by homicide, providing important insights into potential pathways for homicide prevention.[20] Another example was the use of linked data in New Zealand to show the high rate of self-harm in people released from prison, challenges of transitioning from prison to the community and the opportunities to improve the care of people in this situation.[21]

Natural experiments

In recent years, there has been considerable development of methodologies for the evaluation of natural experiments of interventions where the intervention has not been randomised.[22] Natural experiments are an attractive design as they enable evaluations of interventions that are difficult to randomly allocate, such as policy and health system changes, as well as those where the providers find it difficult to randomise their services.[23] Well-designed evaluations using data linkage provide an efficient approach to measuring outcomes in natural experiments.

Recent examples of linkage between service provision data and high-quality research registers provide exemplars for the evaluation of secondary and tertiary prevention initiatives. The Emergency Medicine Retrieval and Transport Service (EMERTS) in Wales provides physicians to emergency events by helicopter and fast cars. An evaluation required severity matched cases who were or were not transported to hospital by EMERTS was achieved through individual record linkage to the Trauma Audit and Research Network database.[24] A 37% reduction in risk-adjusted mortality in patients transported by EMERTS was observed, which resulted in 24-hour, nationwide expansion of the service. Similarly, linked routine Victorian State Trauma Registry and hospital clinical performance and administrative data were used to evaluate the impact of new infrastructure (purpose-built ward) and a new model of allied healthcare and found the new model of care, but not the infrastructure change alone, was highly cost-effective.[25]

Notably, a key limitation of randomised controlled trials can be low external validity. Data linkage can be used to assess whether findings from trials are generalisable to the wider population. For example, the Nurse-Family Partnership in the USA was a randomised trial of nurse home visitations to families at risk of poor health and social outcomes. They found that nurse home visitations reduced ED presentations for injuries and poisoning and resulted in fewer cases of child abuse or neglect.[26] Based on this, similar models were implemented in a number

of jurisdictions. However, the results were not replicated in other areas when population data linkage was used in the evaluation.[27 28] Data linkage helped to demonstrate that the intervention model did not generalise to other settings.

Household data linkage

The ability to link data both at individual household level and at individual levels opens up opportunities to evaluate injury prevention interventions at the household or grouped household level.[29] Housing quality has been shown to be related to a variety of injuries, but studies are few.[30] The Carmarthenshire Housing and Health Study used data linkage to follow 32 009 residents across 8558 social homes to evaluate the health impacts of various housing improvements over 10 years.[31] These authors reported a 39% reduction in emergency admissions in people over 60 overall, along with smaller but still significant reductions in injury admissions associated with some of selected interventions; however the study was not powered to test specific reductions in injuries.[31]

Studies from the UK and New Zealand have used household record linkage to explore the distance between homes and the nearest alcohol outlets as an exposure, with alcohol-related harms as the outcome (alcohol-related hospital attendances and admission and police reported crime).[32 33] These studies found disparate results, which potentially relate to environmental and exposure-related factors across different populations.

Household level linkage would also help augment the quantification of the impact of injuries on cohabiting family members, as recommended in the LOAD framework.[7]

The benefits of NLP in improving data quality and depth for injury prevention research

NLP and machine learning techniques have been applied to unstructured text data for many years in the injury domain, most commonly in the fields of occupational surveillance, ED injury surveillance, product safety surveillance and social media surveillance, particularly in relation to mental-health, self-harm and substance abuse.[34 35]

NLP techniques have been used for: (1) better understanding causal mechanisms involved in injury events by providing more contextual information about direct and underlying mechanisms beyond basic coded categories, (2) capturing pre-event risk factors, mechanisms and object interactions, (3) capturing rare causes/emerging hazards, which would otherwise not have been captured in coded form and (4) automating/semi-automating coding to enable more rapid reporting of data than coding resources allow.[36–38] As LLMs continue to develop, there is significant potential for a flow on effect to enhance NLP techniques to become more sophisticated, accurate and real time to support decision-making.[39] However, such algorithms may make errors and require validation against human expert coding for clinical accuracy.[40] There are also challenges to the use of LLMs due to privacy protection and large computational requirements. Some LLMs require the data to be moved from its original source and fed into the LLM. The difficulty with this is that the text may be highly disclosive of an individual, given that many injury events are reported in the media. An alternative approach is the importation of open source LLMs into the original data source environment or the TRE and the research conducted there. This hampers open science to some extent, but can be ameliorated by sharing of the underlying codes and algorithms. This field of research is promising, but still in its infancy in terms of contributing to injury research.

Closing comments and next steps

As data linkage capabilities evolve and the analytic techniques, which unlock the unstructured data inherent in datasets available for linkage mature, the potential to use these data sources to enhance evidence-informed injury prevention and policy development is evident. We have provided a tangible, but not exhaustive list of examples showing the breadth and value of data linkage, and the emerging capabilities of NLP techniques to enhance injury research. Data linkage capacity continues to grow in high-income, low-income and middle-income countries.

What injury prevention researchers need to do is to carry out a review of the data sources available to them and explore whether there are existing data linkage facilities in their jurisdiction or setting. A good place to start is by searching the IPDLN membership directory for potential collaborators (https://ipdln.org/membership/). Once researchers are informed of the potential for linkage in their area, they can consider what additional questions could be answered through this paradigm. In particular, the embedding of individual or household interventions that are primarily designed to prevent injuries, for example, installation of stair gates or handrails or where injury prevention could be a secondary aim, for example, wealth transfers to reduce overall inequalities in health, into data linkage systems would facilitate the evaluation of interventions that are difficult or impossible to achieve through standard randomised trials.

Another issue is the importance for injury researchers in this area to share their methods, code, models and tools to improve consistence and efficiencies in this field. Increased collaboration between injury prevention researchers and data scientists working on population data linkage systems has much to offer this field of research.

Correction notice This article was updated to CC-BY-NC on 14/11/24.

Contributors RAL contributed to the original conception and design of the paper, prepared the initial draft and undertook revisions and final approval of the paper. BJG and KV contributed to the revised conception and design, contributed to the second draft and revisions and final approval of the paper. RAL is the guarantor.

Funding RAL is supported by the Administrative Data Research Wales grant (ES/W012227/1) funded by the Economic and Social Research Council (ESRC). BJG is supported by a National Health and Medical Research Council of Australia Investigator Grant (L2, ID2009998). KV is supported by funding from the Motor Accident Insurance Commission Queensland.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This work was a combination of expert opinion and literature review and commentary and did not require ethical approval.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Ronan A Lyons http://orcid.org/0000-0001-5225-000X
Kirsten Vallmuur http://orcid.org/0000-0002-3760-0822

REFERENCES

1 Doetsch JN, Kajantie E, Dias V, *et al*. Record linkage as a vital key player for the COVID-19 syndemic - The call for legal harmonization to overcome research challenges. *Int J Popul Data Sci* 2023;8:2131.
2 ADRUK. Understanding data linkage. 2024. Available: https://www.adruk.org/learning-hub/skills-and-resources-to-use-administrative-data/understanding-data-linkage
3 Network PHR. Let's navigate data linkage. 2021. Available: https://www.phrn.org.au/for-researchers/lets-navigate-data-linkage

4 Dowd B, McKenney M, Boneva D, *et al*. Disparities in National Institute of Health trauma research funding: the search for sufficient funding opportunities. *Medicine (Balt)* 2020;99:e19027.
5 Dreyer NA, Mack CD. Tactical Considerations for Designing Real-World Studies: fit-for-Purpose Designs That Bridge Research and Practice. *Pragmat Obs Res* 2023;14:101–10.
6 Kendi S, Macy ML. The Injury Equity Framework - Establishing a Unified Approach for Addressing Inequities. *N Engl J Med* 2023;388:774–6.
7 Lyons RA, Finch CF, McClure R, *et al*. The injury List of All Deficits (LOAD) Framework--conceptualizing the full range of deficits and adverse outcomes following injury and violence. *Int J Inj Contr Saf Promot* 2010;17:145–59.
8 Quistberg DA. Potential of artificial intelligence in injury prevention research and practice. *Inj Prev* 2024;30:89–91.
9 Khurana D, Koli A, Khatter K, *et al*. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* 2023;82:3713–44.
10 Lensen S, Macnair A, Love SB, *et al*. Access to routinely collected health data for clinical trials - review of successful data requests to UK registries. *Trials* 2020;21:398.
11 Kashef MA, Giugliano G. Legacy effect of statins: 20-year follow up of the West of Scotland Coronary Prevention Study (WOSCOPS). *Glob Cardiol Sci Pract* 2016;2016:e201635.
12 Griffiths LJ, Cortina-Borja M, Tingay K, *et al*. Are active children and young people at increased risk of injuries resulting in hospital admission or accident and emergency department attendance? Analysis of linked cohort and electronic hospital records in Wales and Scotland. *PLoS ONE* 2019;14:e0213435.
13 Tingay KS, Bandyopadhyay A, Griffiths L, *et al*. Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort. *Int J Popul Data Sci* 2019;4:579.
14 Boyd A, Thomas R, Hansell AL, *et al*. Data Resource Profile: the ALSPAC birth cohort as a platform to study the relationship of environment and health and social factors. *Int J Epidemiol* 2019;48:1038–9k.
15 Reading R, Jones A, Haynes R, *et al*. Individual factors explain neighbourhood variations in accidents to children under 5 years of age. *Soc Sci Med* 2008;67:915–27.
16 Dipnall JF, Lyons J, Lyons RA, *et al*. Impact of an injury hospital admission on childhood academic performance: a Welsh population-based data linkage study. *Inj Prev* 2024;30:206–15.
17 Mitchell R, Cameron CM, Lystad RP, *et al*. Impact of chronic health conditions and injury on school performance and health outcomes in New South Wales, Australia: a retrospective record linkage study protocol. *BMJ Paediatr Open* 2019;3:e000530.
18 Visnick MJ, Pell JP, Mackay DF, *et al*. Educational and employment outcomes associated with childhood traumatic brain injury in Scotland: a population-based record-linkage cohort study. *PLoS Med* 2023;20:e1004204.
19 Kolaja C, Castañeda SF, Woodruff SI, *et al*. The relative impact of injury and deployment on mental and physical quality of life among military service members. *PLoS ONE* 2022;17:e0274973.
20 O'Neill M, Buajitti E, Donnelly PD, *et al*. Characterising mental health and addictions and assault-related health care use in the year prior to death: a population-based linked cohort study of homicide victims. *Int J Popul Data Sci* 2021;6:1410.
21 Borschmann R, Thomas E, Moran P, *et al*. Self-harm following release from prison: a prospective data linkage study. *Aust N Z J Psychiatry* 2017;51:250–9.
22 Craig P, Katikireddi SV, Leyland A, *et al*. Natural Experiments: an Overview of Methods, Approaches, and Contributions to Public Health Intervention Research. *Annu Rev Public Health* 2017;38:39–56.
23 de Vocht F, Katikireddi SV, McQuire C, *et al*. Conceptualising natural and quasi experiments in public health. *BMC Med Res Methodol* 2021;21:32.
24 Lyons J, Gabbe BJ, Rawlinson D, *et al*. Impact of a physician - critical care practitioner pre-hospital service in Wales on trauma survival: a retrospective analysis of linked registry data. *Anaesthesia* 2021;76:1475–81.
25 Gabbe BJ, Reeder S, Ekegren CL, *et al*. Cost-effectiveness of a purpose-built ward environment and new allied health model of care for major trauma. *J Trauma Acute Care Surg* 2023;94:831–8.
26 Olds DL, Henderson CR Jr, Chamberlin R, *et al*. Preventing child abuse and neglect: a randomized trial of nurse home visitation. *Pediatrics* 1986;78:65–78.
27 Green BL, Sanders MB, Tarte J. Using administrative data to evaluate the effectiveness of the Healthy Families Oregon home visiting program: 2-year impacts on child maltreatment & service utilization. *Child Youth Serv Rev* 2017;75:77–86.
28 Robling M, Lugg-Widger FV, Cannings-John R, *et al*. Nurse-led home-visitation programme for first-time mothers in reducing maltreatment and improving child health and development (BB:2-6): longer-term outcomes from a randomised cohort using data linkage. *BMJ Open* 2022;12:e049960.
29 Lyons RA, Turner S, Lyons J, *et al*. All Wales Injury Surveillance System revised: development of a population-based system to evaluate single-level and multilevel interventions. *Inj Prev* 2016;22:i50–5.
30 DiGuiseppi C, Jacobs DE, Phelan KJ, *et al*. Housing interventions and control of injury-related structural deficiencies: a review of the evidence. *J Public Health Manag Pract* 2010;16:S34–43.
31 Rodgers SE, Bailey R, Johnson R, *et al*. Emergency hospital admissions associated with a non-randomised housing intervention meeting national housing quality standards: a longitudinal data linkage study. *J Epidemiol Community Health* 2018;72:896–903.
32 Connor JL, Kypri K, Bell ML, *et al*. Alcohol outlet density, levels of drinking and alcohol-related harm in New Zealand: a national study. *J Epidemiol Community Health* 2011;65:841–6.
33 Fone D, Morgan J, Fry R, *et al*. Change in alcohol outlet density and alcohol-related harm to population health (CHALICE): a comprehensive record-linked database study in Wales. *Pub Health Res* 2016;4:1–184.
34 Conway M, Hu M, Chapman WW. Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. *Yearb Med Inform* 2019;28:208–17.
35 Vallmuur K. Machine learning approaches to analysing textual injury surveillance data: a systematic review. *Accid Anal Prev* 2015;79:41–9.
36 Catchpoole J, Nanda G, Vallmuur K, *et al*. Application of a Machine Learning-Based Decision Support Tool to Improve an Injury Surveillance System Workflow. *Appl Clin Inform* 2022;13:700–10.
37 Nanda G, Vallmuur K, Lehto M. Intelligent human-machine approaches for assigning groups of injury codes to accident narratives. *Saf Sci* 2020;125:104585.
38 Omaki E, Shields W, Rouhizadeh M, *et al*. Understanding the circumstances of paediatric fall injuries: a machine learning analysis of NEISS narratives. *Inj Prev* 2023;29:384–8.
39 Klang E, García-Elorrio E, Zimlichman E. Revolutionizing patient safety with artificial intelligence: the potential of natural language processing and large language models. *Int J Qual Health Care* 2023;35:mzad049.
40 de Hond A, Leeuwenberg T, Bartels R, *et al*. From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digit Health* 2024;6:e441–3.