Commentary

BIG DATA & SOCIETY

Big Data & Society
October–December: 1–4

DOI: 10.1177/20539517241296058
journals.sagepub.com/home/bds
Sage

With great (statistical) power comes great responsibility: A comment on the ethics of using administrative data to investigate marginalised populations

Louise Marryat[1], Ruchika Gajwani[2], Sharon Graham[2], Marion Henderson[3], Christine Puckering[4], Lucy Thompson[5], Philip Wilson[6] and Helen Minnis[2]

Abstract

As health data infrastructure improves, we have the opportunity to link increasing volumes of data in order to investigate important health problems. This is perhaps most pertinent when looking at the experiences and outcomes of our most disadvantaged groups, who are often invisible in data obtained through primary research. Whilst these data offer enormous opportunity, there are also ethical implications in their use, which are less frequently discussed than in relation to their qualitative counterparts. As a diverse group of clinicians and academics working across public health, we share our experience and understanding of how we can improve our reflexivity in health data science and ensure that research in this area is ethically conducted in co-production with the people whose data we are using. We discuss the potential opportunities, challenges and impacts of using administrative data to investigate marginalised populations.

Keywords

Marginalised populations, ethics, co-production, public engagement, vulnerable populations, drug use

Commentary

From preconception to the grave, vast amounts of data are collected about populations. These include health, education, social care, social security, housing and criminal justice records – 'administrative data'. It is becoming easier to link together and gain access to these data safely and securely. This provides unprecedented opportunities for social scientists and the public to understand life trajectories, and learn more about risks, mediators, moderators and outcomes across a range of fields. Scientists, in partnership with practitioners and those with lived experience, need to use the *right methodologies* to ask the *right questions* of administrative data about health and social care needs, in order to plan services effectively (Astle et al., 2023).

The opportunity for using these administrative data is particularly pertinent when looking at experiences and outcomes for our most disadvantaged groups, often invisible in data obtained through primary research. This can be due to the need to research small sub-groups that are usually underrepresented in research studies, or through difficulties in accessing specific populations for research purposes. For example, Sim et al. attempted to trace and interview 10 mothers with problematic substance use and to assess their children. Seven years after antenatal interview they were unable to interview any: of the 10 children, one was deceased, two had been adopted and one was uncontactable due to the mother living in a women's refuge and the address being secret, three opted out, and the rest remained uncontactable/ unresponsive after each woman received a *minimum* of 10 phone calls and five attempted visits (Sim et al., 2014).

[1]School of Health Sciences, University of Dundee, Dundee, UK
[2]School of Health and Wellbeing, University of Glasgow, Glasgow, UK
[3]School of Social Work and Social Policy, University of Strathclyde, Glasgow, UK
[4]PSPartnership (Scotland) Ltd, UK
[5]Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
[6]Centre for Research and Education in General Practice, University of Copenhagen, Copenhagen, Denmark

Corresponding author:
Louise Marryat, School of Health Sciences, University of Dundee, Dundee, UK.
Email: lmarryat001@dundee.ac.uk

Investigating timelines among populations 'hidden' in traditional datasets

Birth cohort studies generally produce some of the most robust and detailed evidence about many aspects of the life-course. They are not, however, without bias, and these biases can have specific impacts on the quality and quantity of data recorded about people living in the most disadvantaged circumstances. As an example, the Growing Up in Scotland (GUS) study is Scotland's national birth cohort study. The original birth cohort (Birth Cohort 1) comprises around 10% of children born in 2004–2005 in Scotland. Families were first approached for interview when their baby was 10 months old. The original sample came from Child Benefit records, which at the time 97% of families received. After sampling of children was complete, the Department of Work and Pensions (DWP) removed around 5% of the sample: for example, where a child was removed from birth parents or where there had been a death in the family. Children in the most deprived circumstances are more likely to experience these events (Wilkinson and Pickett, 2009). Consequently, sampling is likely to start off with bias excluding some of the most disadvantaged families. Although the response rate in GUS at Sweep 1 was relatively high at 80%, responders were more likely to be older parents, have fewer children and live in rural and less deprived areas (Bradshaw et al., 2007). Such biases are common in cohort studies, and a range of techniques for weighting data based on available information have been developed in order to 'correct' the results for these biases. However, some particularly disadvantaged groups might demonstrate further bias through attrition, for example, families of children with behavioural problems have been found to be more likely to drop out of cohort studies (Wolke et al., 2009), whilst item response bias has been found in domestic abuse data for women who have previously experienced abuse (Skafida et al., 2022).

Administrative data have the potential to circumvent some of these challenges – providing longitudinal insights on individuals who would otherwise be invisible in traditional samples. This was aptly demonstrated by Tweed et al. (2022) who used linked administrative data to investigate the association of premature mortality with one of more of: opioid use, homelessness, psychosis or justice involvement. Through this, they identified 536,653 adults living in Glasgow City Council area, alive and younger than 75 years at the start of follow-up on 1 April 2014. They demonstrated that 5.2% had experienced any one of these exposures, with 1% experiencing more than one. In the following four years, 2.1% of people died, with risk of death far higher for people experiencing any of these problems, and (with the exception of opioid use), having more than one of these exposures reported led to much higher risk of premature mortality. People who used opioids had far higher risk of premature mortality regardless of whether they were recorded as having other experiences or not (Tweed et al., 2022). Similarly linking primary care, hospital activity and mortality data allowed researchers to demonstrate a strong link between missing appointments in family practice and greatly increased morbidity and mortality (McQueenie et al., 2019). Arguably, only scientists with an interest in the topic area (as opposed to analysts purely working within health services asking routine questions of these data) would have thought to combine the variables and use the data in that way.

Ethics of and scientific justification for collecting population data

To quote Spider-man, 'with great power comes great responsibility'. In this case, we have a responsibility towards those who have provided their data and those potentially affected by the findings. We need to think carefully about *how* routine data are used, and the ethical implications. Data access is usually carefully scrutinised – in Scotland this is through the Public Benefit and Privacy Panel. The panel includes four members of the public (out of a total of 15), however the extent to which those members are from the most disadvantaged groups is unclear, particularly in terms of representing often excluded groups (i.e. those who have used drugs or are care experienced). A Freedom of Information request (ref: 2022-001496) revealed that Public Health Scotland did not hold data on the numbers of panel members by deprivation quintile, or with care experience. In research, we know that the most disadvantaged groups are likely to be underrepresented (Langer et al., 2021): this is equally likely to apply in mainstream patient and public involvement in science oversight. Although we are required to state our public and patient involvement when seeking to access administrative data, the extent to which this is scrutinised varies. Research on peoples' views of their data being shared for research suggests that while a minority, mainly from ethnic and religious minorities and people from more disadvantaged communities, have been less likely to provide support for data sharing, an overall majority (even from these disadvantaged communities) *are* in favour (Jones et al., 2022). Nevertheless, concerns about data accessibility have led to unwieldy governance procedures that can slow or prevent the answering of important research questions and are often disproportionate to the risk (Emery-Barker et al., 2008). 'The Promise', the main output of the Scottish Independent Care Review (into the care of children looked after by local authorities), was a good example of a process driven by care-experienced people that concluded proper use of data is imperative and requires involvement of 'experts by experience' in the shaping and interpreting of data: 'Care experienced children and young adults must have ownership over their own stories and personal data so that they can understand and influence how their stories are

shared' (Independent Care Review, 2020: 32). For heavily stigmatised groups, such as those who use drugs, there is currently no evidence about their views on data sharing. It is reasonable to think that such groups may be particularly cautious about sharing of what could be deemed to be extremely sensitive data. Minority groups may also hold concerns about reidentification relating to small numbers, although processes are usually robust to prevent this (Harron et al., 2017). Views on data sharing may additionally be culturally rooted: in the Nordic countries, the sharing of personal data is generally viewed as a citizen's duty (Belfrage et al., 2021), whilst the Care.data scandal in England (a failed attempt to routinely link health and social care data after 1 million people opted out due to concerns about commercial data sharing) has resulted in a mistrust of data linkage (Carter et al., 2015).

Involving stigmatised groups in shaping the way administrative data are used could have important implications for science. Quantitative science needs to address 'reflexivity' in the asking and interpretation of research questions, in similarity to good practice in qualitative research. Reflexivity is defined as 'the researchers' engagement of continuous examination and explanation of how they have influenced a research project from choosing a research question to sampling, data collection, analysis and interpretation of data' (Williams et al., 2020). It is currently rare to see this addressed in a quantitative paper. This is at least partly due to assumptions around the objectivity of quantitative research. Quantitative research, including the use of administrative data is, however, influenced by who collects data, what is routinely collected, what questions are asked of these data and how they are analysed. An example is reflected in analyses of obesity data in children, in which extremely high levels of obesity were found at the 27–30-month universal health visitor check. This did not reflect other data on the preschool population. The researchers questioned whether health visitors were selectively weighing heavier children, despite this measurement in theory being a routinely collected population level data item (Horne et al., 2022). Jamieson et al. (2022) have recently published a guide to reflexivity in quantitative research, demonstrating moves towards acknowledging this in the quantitative field.

A stronger input from disadvantaged groups might alleviate some of these concerns. But how best do we involve experts by experience in shaping the research questions needing to be asked of administrative data? Reflections from involving disadvantaged or stigmatised populations in research suggest strategies such as including partnering with community groups, approaching potential participants where they are, using population-appropriate modes of communication, conducting study activities in familiar settings and at convenient times, maintaining frequent contact and offering payment for their time (Langer et al., 2021). Ethically, as researchers we much be aware of the potential for upset and re-traumatising of discussions around this type of sensitive research, and ensure appropriate support is made available for experts by experience, as we would for research participants. Amongst the authors, we have conducted some qualitative research with a particularly stigmatised group – parents whose children have recently been taken into foster care – and found that although sometimes participants reported not fully understanding the research at the time of giving consent, they generally were happy that they had taken part and were contributing to learning about families like themselves. Genuine involvement of experts by experience is an evolving practice, and funding bodies, such as the National Institute for Health and Care Research in the UK, are increasingly supportive of, for example, expert-by-experience co-investigators who are fully involved in shaping the questions and driving the research from beginning to end (Involve, 2005). We have recent experience of this in the Partnership for Change project in which a new infant mental health intervention is being codesigned by experts by experience, researchers and clinicians. We have learned that it is crucial to keep information about the research process simple and understandable, and to pace the research timetable over longer periods in order to allow genuine participation of people who are leading very complex and challenging lives. This is crucial if expert by experience voices are to be heard, if the messages are to be trusted and acted upon, and if experts by experience are to be involved in a way that is not simply tokenistic. This will eventually mean genuine involvement of expert by experiences (including from disadvantaged and stigmatised groups) at all the levels of the research process, including at the most senior level of research governance and policy.

In conclusion, administrative data are an important resource through which crucial questions about health and social outcomes can be asked about some of the most disadvantaged members of the population. This could enable new interventions to be developed and existing ones to be better targeted at those who would benefit most. To ensure the collection and use of administrative data truly serves these disadvantaged members of the population, it will be crucial that we learn how best to involve the people to whom the research findings will be of most concern in a meaningful way. This involvement should start from the development of the research questions, through decisions about data collection, and on through the entire research process. Methods used in qualitative research and public engagement, including interviewing, group discussions and creative methods, can be usefully employed to support this work. Genuine involvement of those most likely to be affected by research findings might also allay fears about potential harms from using administrative data. This could free us up to ask important questions more quickly, so that data can be of most use for service and policy development, benefitting those with the greatest need.

Acknowledgements

The authors would like to acknowledge the input to discussions on this theme, which helped shape our thinking, by Professor Anthony Pelosi.

Dedication

This paper is dedicated to our friend, colleague, and long-term Epi Club member, Professor James Law, who did more than any other scientist to clarify the lifelong significance of early language problems.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Louise Marryat is funded by UKRI ESRC as part of a New Investigator Fellowship. Grant number ES/T015721/1.

ORCID iDs

Louise Marryat https://orcid.org/0000-0002-6093-4679
Marion Henderson https://orcid.org/0000-0001-7582-9516
Lucy Thompson https://orcid.org/0000-0001-7461-3262

References

Astle DE, Moore A, Marryat L, et al. (2023) We need timely access to mental health data: Implications of the Goldacre review. *The Lancet Psychiatry* 10(4): 242–244. DOI: 10.1016/S2215-0366(23)00030-5.

Belfrage S, Lynöe N and Helgesson G (2021) Willingness to share yet maintain influence: A cross-sectional study on attitudes in Sweden to the use of electronic health data. *Public Health Ethics* 14(1): 23–34. DOI: 10.1093/phe/phaa035.

Bradshaw P, Tipping S, Marryat L, et al. (2007) Growing up in Scotland sweep 1–2005 user guide. *Scottish Executive*. 1–12.

Carter P, Laurie GT and Dixon-Woods M (2015) The social licence for research: Why care. Data ran into trouble. *Journal of Medical Ethics* 41(5): 404–409. DOI: 10.1136/medethics-2014-102374.

Emery-Barker J, McClure I, Wood A, et al. (2008) Bypassing bureaucracy to answer important questions quickly. *Journal of the Royal Society of Medicine* 101(5): 217–218. DOI: 10.1258/jrsm.2008.08004.

Harron K, Dibben C, Boyd J, et al. (2017) Challenges in administrative data linkage for research. *Big Data & Society* 4(2). DOI: 10.1177/2053951717745678.

Horne M., Marryat L. and Wood R. (2022) *Universal Health Visiting Pathway evaluation: Phase 1 report - routine data analysis - baseline outcomes.* Scottish Government.

Independent Care Review (2020) The Promise. https://thepromise.scot/resources/2020/the-promise.pdf.

Involve (2005) People and Participation: How to put citizens at the heart of decision making. https://involve.org.uk/sites/default/files/field/attachemnt/People-and-Participation.pdf.

Jamieson MK, Govaart GH and Pownall M (2022) Reflexivity in quantitative research: A rationale and beginner's guide. *Social and Personality Psychology Compass* 17: e12735. DOI: 10.1111/spc3.12735.

Jones R. D., Krenz C., Griffith K. A., et al. (2022) Patient experiences, trust, and preferences for health data sharing. *JCO Oncology Practice* 18(3): e339–e350.

Langer SL, Castro F, Chen G, et al. (2021) Recruitment and retention of underrepresented and vulnerable populations to research. *Public Health Nursing* 38(6): 1102–1115. DOI: 10.1111/phn.12943.

McQueenie R., Ellis D. A., McConnachie A., et al. (2019) Morbidity, mortality and missed appointments in healthcare: a national retrospective data linkage study. *BMC Medicine,* 17(1), 2. DOI: 10.1186/s12916-018-1234-0.

Sim F., Pritchett R., Hepburn M., et al. (2014) Invisible children: Attempting to engage the most vulnerable families. *The British Journal of Psychiatry* 205(2): 158.

Skafida V, Morrison F and Devaney J (2022) Answer refused: Exploring how item non-response on domestic abuse questions in a social survey affects analysis. *In Survey Research Methods* 16(2): 227–240. DOI: 10.18148/srm/2022.v16i2.7823.

Tweed EJ, Leyland AH, Morrison D, et al. (2022) Premature mortality in people affected by co-occurring homelessness, justice involvement, opioid dependence, and psychosis: A retrospective cohort study using linked administrative data. *The Lancet Public Health* 7(9): e733–e743. DOI: 10.1016/S2468-2667(22)00159-1.

Wilkinson R and Pickett K (2009) *The Spirit Level: Why More Equal Societies almost Always do Better*. London: Allen Lane.

Williams V, Boylan A-M and Nunan D (2020) Critical appraisal of qualitative research: Necessity, partialities and the issue of bias. *BMJ Evidence-Based Medicine* 25: 9–11. DOI: 10.1136/bmjebm-2018-111132.

Wolke D, Waylen A, Samara M, et al. (2009) Selective drop-out in longitudinal studies and non-biased prediction of behaviour disorders. *British Journal of Psychiatry* 195(3): 249–256. DOI: 10.1192/bjp.bp.108.053751.