Combining large linked social service microdata and geospatial data to identify vulnerable populations in New Zealand

Lukas Marek[1], James Greenwell[3], Matthew Hobbs[1], John McCarthy[3], Jesse Wiki[1], Malcolm Campbell[1,2], Simon Kingham[1,2], Melanie Tomintz[1]

[1]GeoHealth Laboratory, Geospatial Research Institute, University of Canterbury, Christchurch, New Zealand
[2]School of Earth and Environment, College of Science, University of Canterbury, Christchurch, New Zealand
[3]Ministry of Health, Health and Disability Intelligence, Wellington, New Zealand

Abstract
New Zealand has a large research database called the Integrated Data Infrastructure (IDI). The IDI contains linked microdata about people and households from a wide range of authoritative datasets and surveys originating in and outside government. The GeoHealth Laboratory at the University of Canterbury in collaboration with the New Zealand Ministry of Health supports policymakers and stakeholders across the New Zealand health system through several IDI research projects. One such project aims to determine how an individual's place of residence and the frequency of residential movement can affect access to regionally-provided health and community services. The presented case study examines the relationships between limited or irregular interaction with health services and potential drivers such as low-engagement, non-enrolment or transience during the study period (August 2013 – July 2018).

Keywords: Integrated Data Infrastructure (IDI), health services utilisation, vulnerable populations, transience, New Zealand

1. Introduction

New Zealand has several large research databases containing microdata about people and households from a wide range of authoritative datasets and surveys originating in government agencies and non-government organisations. There is the Longitudinal Business Database (LBD) containing microdata about businesses and the Integrated Data Infrastructure (IDI) that focuses on people visiting or living in New Zealand. Moreover, records in both data sources can be linked. As a result, approved IDI researchers from a range of institutions can access data on topics including business, education, income and work, social services, health, housing, justice, population and communities. Currently the IDI database holds over 166 billion records, is continuously growing and is updated every fourth months (Stats NZ 2017). The IDI spine dataset, the core of the IDI, contains data for over 9 million individuals who have ever lived in New Zealand as well as tens of millions of visitors (Stats NZ 2016). The IDI also contains the locational identifier such as area identifier that allows spatial and regional analyses of the data.

To grant access to the IDI, Stats NZ created a process that evaluates project applications based on the project's contribution to the public good, credibility of the project team and aim to make research publicly available. Researchers can then access the IDI, once their project is approved and researchers absolved the induction. As the IDI holds confidential data, one has to use Stats NZ's secure virtual environment in approved facilities that are located in Stats NZ spaces or other designated facilities (e.g. universities) around New Zealand. All necessary software for data extraction, manipulation and analysis is provided in the closed environment managed by Stats NZ, the hardware is provided by the facility (Stats NZ, universities, government agencies). The GeoHealth Laboratory at the University of Canterbury, in collaboration with the New Zealand Ministry of Health, supports policymakers and stakeholders across the New Zealand health system through several IDI research projects. One project aims to determine how an individual's place of residence and the frequency of residential movement affects access to regionally-provided health and community services. The project also examines associations between limited or irregular interaction with health services and potential drivers such as low-engagement, non-enrolment or transience (Marek, Tomintz, and Campbell 2019).

Through understanding how the frequency of residential movement and health service enrolment are related, effective and targeted health and related social services can be developed (Marek, Tomintz, and Campbell 2019). More importantly, linked data only available in the IDI can be used to enhance cross-sector service integration to improve equity of health service access for some of the most vulnerable populations in New Zealand. The large linked dataset used in this project can provide more robust population estimates, and extract additional value from existing administrative data.

Linked data in health-related research

Globally, there has been a broad and growing interest in the potential for big or linked data to enhance our understanding of a wide variety of topics in medicine and public health (Timmins et al. 2018), including mental health (Stewart and Davis 2016), obesity (Vogel et al. 2019), and broader healthcare (Raghupathi and Raghupathi 2014).

The use of linked data represents a significant development in public health research. As outlined by a recent systematic review (Young and Flack 2018) using big or linked data allows researchers to assess health outcomes on specific cohorts, such as Indigenous populations (Katzenellenbogen et al. 2011, 2013) or populations who are difficult to recruit (McCallum et al. 2014). The linkage of different datasets also represents a platform for longitudinal study, epidemiological surveillance on rare disease (Gibson, Jorm, and McIntyre 2015), and the capture and comparison of healthcare costs across populations and over time (Young and Flack 2018). Consideration should also be given to the limitations of using such data however, with one major pitfall to progress in this field being the management and sharing of data (Mittelstadt and Floridi 2016). Other barriers include the acquisition and storage of such data for research purposes (Vogel et al. 2019).

The development of data linkage methods to provide researchers access to linked data while preserving privacy is not a new phenomenon. For instance, Western Australia (WA) was the first Australian state to initiate data linkage in the 1970s but its use was limited in scope and purpose (Hobbs and McCall 1970). In New Zealand, it has been possible to link all official health datasets using a unique identifier, the National Health Index (NHI) number, since 1993. However, it was previously much more difficult to link health data with data from other sources (Atkinson and Blakely 2017), including geospatial data. Studies often involved ad hoc data linkage, with different versions of the same data being stored in multiple sites, resulting in issues of data integrity and security (Atkinson and Blakely 2017). The availability and use of linked data in public health has grown rapidly in recent years, however, largely facilitated by significant improvements in computing hardware and software (Atkinson and Blakely 2017).

There has also been an increase in the use of data that is repurposed for research purposes. For example, Connelly et al. (2016) make a helpful distinction between data that are 'made' and data which are 'found'. 'Made' data include information collected to investigate a defined hypotheses. In contrast, 'found' data are collected for alternative (non-research) purposes. Examples of 'found' data include online activities such as web searches, commercial transactions such as food purchases, and environmental sensors such as GPS data. The use of novel data, including 'found' data not initially collected for public health research, provides valuable information and can potentially improve understandings of the interaction between, and relative influence of the various determinants of health (Vogel et al. 2019).

Investment in data linkage infrastructure in Australia (Young and Flack 2018) and New Zealand (Stats NZ 2019a) has resulted in a steady increase in the number of research publications involving the use of linked health and human services data. A recent review showed that from 2009 to 2017, a nearly 3.5-fold increase in the number of publications using linked data was observed (Young and Flack 2018). In addition, the review identified the increasing importance of cross-sectoral linkages of health data with education, justice, police, child protection and environmental data (Young and Flack 2018).

Similarly, New Zealand researchers are currently using the IDI to study all-of-government service utilisation, develop accurate population estimates and projections, and understand population transience. This study adapted methods developed from a previous study of population transience (Jiang, Pacheco, and Dasgupta 2018) in order to understand associations between population transience, health service interactions, health service accessibility, family harm, child abuse, and the sociodemographic characteristics of those in the Lakes District Health Board (DHB) catchment (Marek, Tomintz, and Campbell 2019).

Background of the study

This study aims to determine how current residential location and frequency of change affects enrolment for health services, service access, and health outcomes. The New Zealand Public Health Act 2000 established District health Boards (DHBs, of which there are currently 20) to provide health and disability services to populations within geographic areas. Their role is to promote inclusion and participation in society and independence of people with disabilities and to reduce, and eventually eliminate, health disparities. In addition to funding public hospital services, DHBs provide funding to 31 Public Health Organisations (PHOs) which are responsible for ensuring the provision of essential primary health care services, mostly through general practices to those who are enrolled with the PHO. Both DHBs and PHOs have a responsibility to improve and maintain the health of their populations.

This project focuses on the Lakes DHB catchment which serves a population of approximately 110,000 as of 2018. The area was selected due to the interest of local stakeholders. The Lakes DHB is located in the central North Island around lake Taupō in the south and lakes Rotorua, Rotoiti, and Tarawera (and many others) in the north. The most populated places in the area are Rotorua (population 72,500) and Taupō (population 24,700). The Lakes DHB population consists of 50.8% Europeans, 39.3% Māori, 5.0% Asians, 2.6% Pacific People, 1.4% Other ethnicities and 0.8% MELAA. Age-wise, there are 23.7% of children and youth under 17 years of age, 26.5% of the population is 17–39 years old, 37.5% is 40–69 years old, and 12.2% is older than 69 years of age.

What are the current challenges?

DHBs and PHOs largely operate as alliances within geographic boundaries at a DHB level. As part of their responsibility to improve the health of their populations, these alliances organise and subsidise primary care and pharmaceuticals based on population size and other socioeconomic characteristics. They must also integrate health services into a seamless continuum of care.

The estimation of PHO enrolments is crucial in the provisioning and planning of health services in New Zealand. A census-derived population, commonly used in the estimation of enrolment rates, does not account for the dynamic nature of people who move from place to place within and outside a DHB catchment.

Alternatively, IDI-sourced population estimates are based on population level microdata updated three times a year. Therefore, population counts, PHO enrolments and health service utilisation data sourced from the IDI is more current than census-based estimates.

Research questions

This study focused on population transience and health service utilisation during the period August 2013 to July 2018) in Lakes DHB. We also analysed how population transience may be associated with demographic patterns. The aim was to demonstrate how the IDI can be used to link administrative health data with other social sector data in order to provide policy insights about the upstream socioeconomic determinants of health. Specific research questions developed in consultation with health system stakeholders included:

- What are the levels of population transience in Lakes DHB and how do they compare to New Zealand as a whole?
- What are the associations between transience and health service engagement/enrolment?

2. Methods

For the purpose of the study, we considered a variety of datasets within the IDI to identify whether an individual has been in contact with health-related services including pharmacies, laboratory claims, PHO enrolment, hospital discharges, and others. We then estimated the proportion of the Lakes DHB population not enrolled with a PHO or who have not had any contact with health services in the last three years of the reference period. We then compared enrolment distribution with the transience level assigned to everyone, the IDI data sets (family-related harm, child abuse, age, ethnicity, and chronic conditions) and datasets uploaded ad hoc to the IDI such as socioeconomic deprivation, access to health services.

The IDI datasets used for this study were:

- Population - ethnicity, age, area level socioeconomic deprivation, derived transience;
- PHO enrolments;
- National non-admitted patient collection;
- Public hospital discharges;
- Private hospital discharges;
- Chronic conditions;
- Pharmaceuticals;
- ACC (Accident Compensation Corporation) claims (injuries).

Sample population identification

This study followed the methodology developed for the Ministry of Health funded SUPERU report (Section 3: Definitions and populations of interest) (Jiang, Pacheco, and Dasgupta 2018) to create the sample population, applying several key modifications. The address table registering the change of residential address for individuals living in New Zealand served as a basis. Records were filtered from a change of residential address table (during reference period August 2013 to July 2018) based on whether individuals were present in linked data tables from various IDI sources. The following set of instructions describes the modified filtering procedure:

- Only records relevant to the reference period (August 2013 to July 2018) from the address change table were retained;
- individuals with death records in the dataset from the Department of Internal Affairs were removed;
- individuals with death records from other sources were removed (Ministry of Health – mortality and population cohort);
- all children born in the reference period were included;

- individuals who held a New Zealand residence class visa and lived in New Zealand for at least half of the reference period were included;
- individuals who either left New Zealand before the reference period or spent less than 50% of the reference period in New Zealand were removed using the overseas spell table;
- individuals not existent in the IDI spine (filtered based on the birth year and missing records) were removed;
- individuals with missing address and/or meshblock (smallest geographic unit for which statistical data is reported (Stats NZ 2019b)) information were removed;
- the address change records were collated if consecutive spells shared the same address;
- individuals with missing deprivation information were removed.

A score from the area-level New Zealand Index of Socioeconomic Deprivation (NZDep2013) was appended to each case in the subset population, producing a sample population dataset. This population was then supplemented with personal details from the IDI spine. The final population data contained frequency of residential address change, all known residential addresses based on government service utilisation, and matching socioeconomic deprivation and personal demographic information (gender, age, and ethnicity). The subset of this population residing within Lakes DHB was then based on relevant meshblock and area code.

Assigning the level of transience

A transience category was assigned to the sample population. Categorical thresholds defined by previous research (Jiang, Pacheco, and Dasgupta 2018) were retained ensuring comparability with existing research, but the category assignment rules were adjusted to comply with a longer reference period (August 2013 to July 2018). The categories were named *Non-movement* (No address change during the reference period), *Low movement* (1–2 address changes), *Medium movement* (3–4 address changes), and High movement (five or more address changes). The latter was further broken into *Vulnerable transient,* *Transient*, and *High movement (upward)* based on the trajectory of socioeconomic deprivation related to former and current residential address.

The *Vulnerable transient* category was assigned to individuals who either moved towards high deprivation areas (NZDep2013 Index 8–10), moved exclusively within high deprivation areas, or moved ten times or more (those moving at least twice a year on average) within the reference period regardless of deprivation. Individuals were categorised as *Transient* when they moved within medium deprived areas (NZDep2013 Index 4–7) or moved from a low deprivation area (NZDep2013 Index 1–3) to a medium deprivation area. The *High movement (upward)* category contains individuals who either moved within areas of low deprivation (NZDep2013 Index 1–3) or moved from areas of high/moderate deprivation to areas of low deprivation.

Interaction of the Lakes DHB population with healthcare services and the population transience

Following transience assignment for the population of Lakes DHB we compared population subsets with other datasets in the IDI related to healthcare interaction. We also attached a derived accessibility index to each individual. The Lakes DHB population was categorised into three groups regarding PHO enrolments. The first group were those individuals who *did not have a record in the PHO enrolments database*; members of the second group *were enrolled in a PHO, but did not consult with a PHO in the last three years* of the reference period; and the third group were people who *enrolled in a PHO and who consulted with a PHO* in the last three years.

3. Results

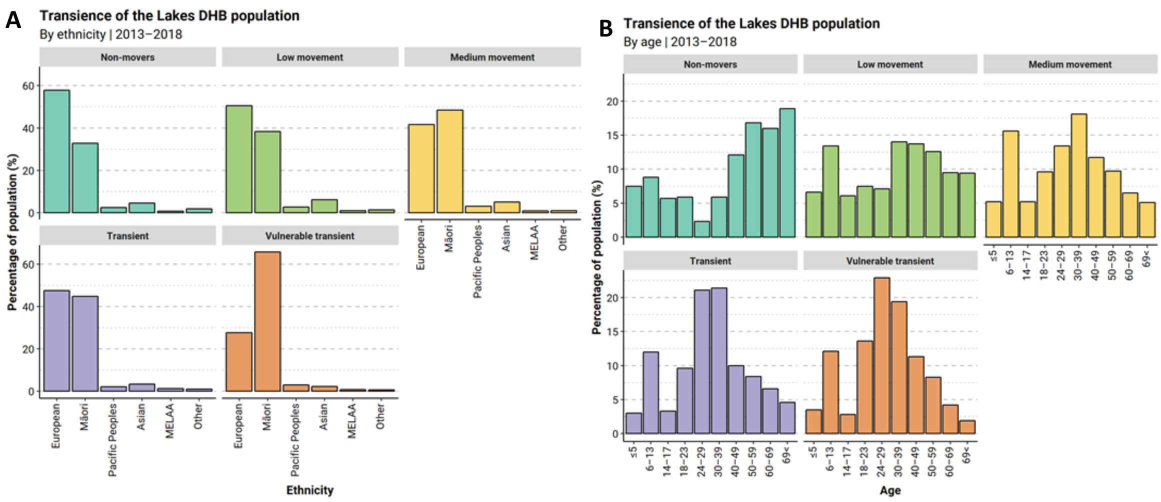
Population transience in New Zealand and Lakes District Health Board

We identified the sample New Zealand population, comprising of approximately 4.63 million people, for the reference period August 2013 to July 2018. Table 1 shows the transience of the New Zealand population and the Lakes DHB population within the reference period of the study. There is a higher proportion of the *Vulnerable Transient* population (6.4%), a higher proportion of Medium-movers (14.4%), and a lower proportion of Non-movers (42.0%) living in Lakes DHB compared to the New Zealand average.

Tab. 1 Comparison of the level of the transience of the New Zealand population and Lakes DHB population within the reference period August 2013–July 2018.

	New Zealand	Lakes DHB
Non-movers	45.9%	42.0%
Low movement	35.8%	35.3%
Medium movement	12.6%	14.4%
High movement upward	0.3%	0.1%
Transient	1.8%	1.7%
Vulnerable transient	3.7%	6.4%
Population	4,634,634	111,120

Figure 1a and Figure 1b demonstrate the ethnicity and age structure of the population in Lakes DHB, by assigned transience category. As shown, there is a significantly higher proportion of Māori within the *Vulnerable transient* category (65.7%). Additionally, it is noticeable that people between 24 and 38 years of age tend to move more, as they form almost half of the population structure of *Transient* and *Vulnerable transient* groups.



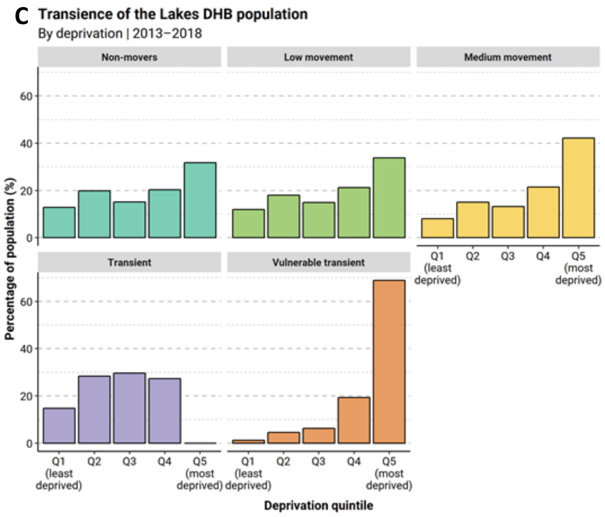


Fig. 1 Transience of Lakes DHB population by (a) ethnicity, (b) age structure, and (c) deprivation

Figure 1c shows that up to 35.9% of the population in Lakes DHB live in the most socioeconomically deprived areas (Q5) and 20.8% in Q4, while only 11.2% of people live in the least deprived areas (Q1). There is a considerable proportion of the population living in the most deprived areas in all transience categories, except for the *Transient* category. We can see that even 31.8% of *Non-movers* and 33.8% of *Low-movers* live in the most deprived areas of Lakes DHB long-term.

Lakes DHB population and PHO enrolments

The majority of the Lakes DHB population (86.9%) is enrolled in a PHO and also consulted with a General Practitioner (GP) in the last three years of the reference period. However, a considerable percentage of people did not enrol in a PHO (5.9%) or enrolled but did not consult recently (7.2%). Also there is a small group of the population (3.7%) that interacted with health-related services, although not being enrolled.

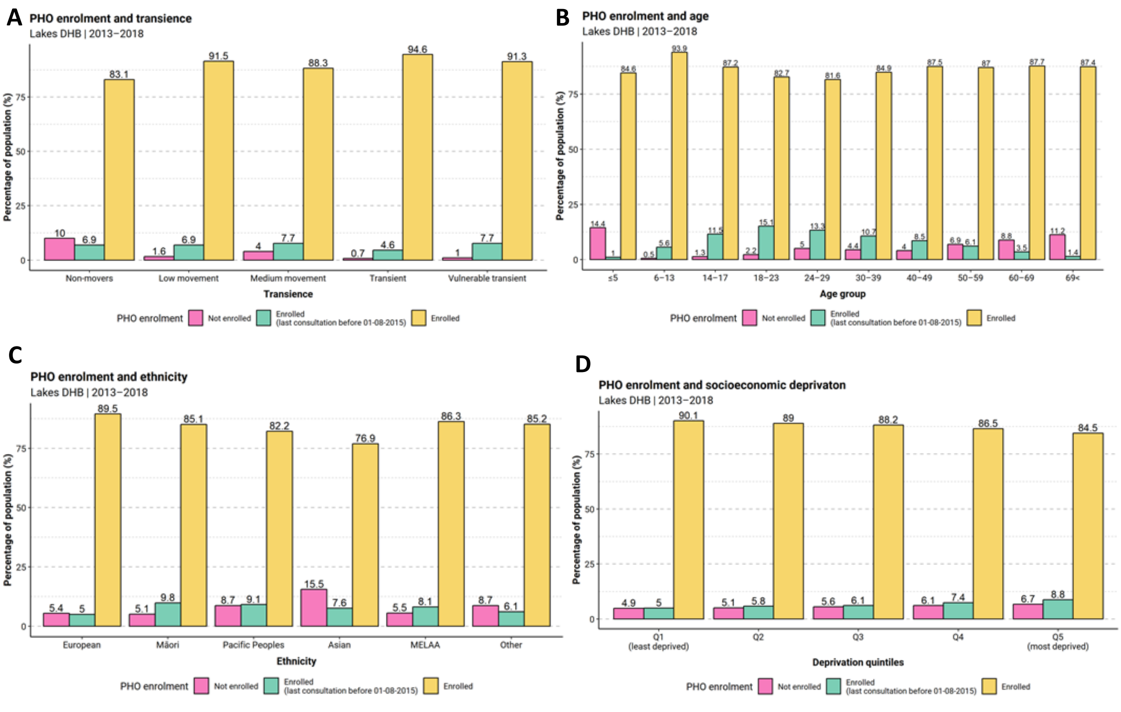


Fig. 2 PHO enrolment of the Lakes DHB population by (a) transience, (b) age, (c) ethnicity, and (d) deprivation

Generally, the more mobile (due to residential address changes) a person in Lakes DHB is, the more likely they are to be enrolled in a PHO. Figure 2a shows that the highest proportion of the unenrolled population is among *Non-movers* (10%) and *Low-movers* (4%), while this group only comprises 1% or less of the *Vulnerable transient* and *Transient* populations. The percentage of the population enrolled but consulting in the last three years is around 7% in all transience categories, except the *Transient* group (4.6%). *Non-movers* have the highest number of people who interact with health services but are not enrolled in a PHO (5.8%).

The highest proportion of the unenrolled population in Lakes DHB was among the group of children under five years of age (Fig. 2b). However, this group also has the highest proportion of people not enrolled who interacted with health care services. The percentage of unenrolled people is lowest among children aged 6–13 and then rises again with age, reaching another high peak in the oldest population (above 69 years). Interaction with a PHO drops from birth to 18–23 years of age, then rises with age (Fig. 2b).

Regarding ethnicity and enrolment, the highest proportion of those not enrolled are of Asian ethnicity (15.5%), while Māori (5.1%) and Europeans (5.4%) have higher enrolment (Fig. 2c). On the other hand, Asian people were the group that interacted with health services the most without being enrolled (9.0%). Although the enrolment of Māori is high in Lakes DHB, almost 10% did not interact with a PHO in the last three years of the reference period (Fig. 2c).

We identified a mild gradient in the percentage of those not enrolled and the proportion of the population that enrolled in a PHO but who did not consult with a GP in the last three years of the reference period, both of which were associated with the socioeconomic deprivation of the population (Fig. 2d). This suggests an increasing proportion of the population that does not interact with health services is associated with increases in socioeconomic deprivation.

4. Discussion

The IDI is an invaluable resource that enables researchers to access and link multiple authoritative administrative databases. However, the results of analyses still require critical evaluation as the data are recorded by many agencies with varying levels of data quality. For instance, IDI records for an individual are based on the direct interaction of the individual with government agencies. Therefore, those not interacting with the government (on a regular basis) can be easily missed by IDI analyses just as those who do not regularly consult with a GP eventually disappear from the PHO enrolment register (usually after 3 years and attempt to contact the patient). Another potential cause for underrepresentation could be the wide-spread use of prioritised ethnicity (as opposed to self-defined ethnicity or multiple ethnicities) (Cormack and Robson 2010; Leather 2009). Prioritised ethnicity is commonly used for research purposes due to simplicity but it can under- or overestimate the real number of people of certain ethnicities.

The sample New Zealand population derived in this case study is bigger than the population derived in the SUPERU report ((Jiang, Pacheco, and Dasgupta 2018) due to multiple reasons. Firstly, several filtering queries were altered in a way that conditions related to time spent in New Zealand corresponded throughout the population. This applied to the 'rule of 50%' used in the selection of individuals living abroad (at least 50% of time spent in New Zealand). Consequently, the residency condition and the youngest children conditions changed accordingly. Secondly, there can be syntactical differences in SQL queries used in the process of data filtering and extraction. Thirdly, there can be slight differences in the versions of IDI and data used. Nevertheless, there were major dissimilarities identified during the filtering of children, people with an overseas spell, and visa applicants.

Censoring due to small numbers is another factor of this research which may distort actual population size. We were required to aggregate small population subsets to ensure privacy protection and confidentiality of the data.

The *High movement (upward)* group was often aggregated with the *Transient* group in order to protect confidentiality. Due to this, some of the graphical outputs have a reduced number of groups when compared to complete categorisation. Privacy and confidentiality protection necessitated a significant degree of censoring for this project.

The goal of the IDI is not only to share the otherwise inaccessible data sets with a wider range of researchers but also to foster collaboration among researchers (Stats NZ 2019c). As health-related research is the most common topic investigated using Stats NZ microdata (Stats NZ 2019a), the research group called Virtual Health Information Network (VHIN) has been formed to support collaboration by actively developing a network of researchers and sharing IDI best practices and potential pitfalls.

5. Conclusions

This study combines and analyses multiple datasets within the IDI using a variety of statistical and mapping methods. The degree of transience in the population of New Zealand was estimated with a particular focus on Lakes DHB. The analysis identified that transience is positively associated with PHO enrolment and health services interaction, meaning the majority of people who were not enrolled in a PHO (or did not interact with the health system in the last three years) are people who do not change their residential address frequently.

The study is based mostly on the descriptive statistics of multiple linked IDI data sources. While data visualisation can provide insight into the spatial variations of the population within Lakes DHB, we produced minimal map outputs due to data handling protocols for confidentiality. We identified that a higher proportion of the *Vulnerable transient* population lives in Lakes DHB (6.4%) compared to the New Zealand average (3.7%). Also, the more people in Lakes DHB are mobile, the more likely it is that they are enrolled in a PHO and that they recently interacted with health services. Furthermore, we found that the highest proportion of the unenrolled population in Lakes DHB was found in the group of children under five years of age.

Results of the study provide an evidence base that may be useful for health and disability sector stakeholders in order to improve their understanding of the population behaviour and for the design of new screening and support interventions. Evidence underpinning health policies if often derived from health system data that may include limited demographic information. Large linked social sector datasets such as the IDI enable socio economic determinants of health to be quantified at a breadth and scale that has not existed before in New Zealand. This study goes beyond using basic demographics and siloed health sector data to understand health outcome inequities between population groups. It is too soon to discuss how health authorities will be using information from this project, however, it is clear the new insights have raised awareness of this research potential and raised new more targeted questions from health sector leaders.

Acknowledgement

This research was carried out as part of the GeoHealth Laboratory work programme at the University of Canterbury, funded by the New Zealand Ministry of Health. The authors would also like to thank stakeholders from Lakes District Health Board and Rotorua Area Primary Health Services as well as the authors of SUPERU Transience study. We thank the IDI team at Statistics NZ for their input and use of data.

Disclaimer

The results in this report are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) managed by Statistics NZ. The opinions, findings, recommendations and conclusions expressed in this report are those of the author(s), not Statistics NZ or other government agencies.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business or organisation. The results in this report have been made confidential to protect these groups from identification.

Careful consideration has been given to the privacy, security and confidentiality issues associated with using administrative and survey data in the IDI. Further details can be found in the privacy impact assessment for the IDI available from www.stats.govt.nz.

References

Atkinson, June, and Tony Blakely. 2017. "New Zealand's Integrated Data Infrastructure (IDI): Value to Date and Future Opportunities." International Journal of Population Data Science 1 (1): 23889. https://doi.org/10.23889/ijpds.v1i1.124.

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." Social Science Research 59: 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015.

Cormack, Donna, and Carey Robson. 2010. "Classification and Output of Multiple Ethnicities: Considerations for Maori Health." Wellington. https://www.fmhs.auckland.ac.nz/assets/fmhs/Te Kupenga Hauora Māori/docs/classification.pdf.

Gibb, S, Bycroft, C, Matheson-Dunning, N (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI). http://archive.stats.govt.nz/~/media/Statistics/surveys-and-methods/methods/research-papers/topss/identifying-nz-resident-pop-in-idi/identifying-nz-resident-population-in-idi.pdf

Gibson, Amy, Louisa Jorm, and Peter McIntyre. 2015. "Using Linked Birth, Notification, Hospital and Mortality Data to Examine False-Positive Meningococcal Disease Reporting and Adjust Disease Incidence Estimates for Children in New South Wales, Australia." Epidemiology and Infection 143 (12): 2570–79. https://doi.org/10.1017/S0950268814003355.

Hobbs, M.S.T., and M.G. McCall. 1970. "Health Statistics and Record Linkage in Australia." Journal of Chronic Diseases 23 (5–6): 375–81. https://doi.org/10.1016/0021-9681(70)90020-2.

Jiang, Nan, Gail Pacheco, and Kabir Dasgupta. 2018. "Residential Movement within New Zealand: Quantifying and Characterising the Transient Population." Auckland. http://www.superu.govt.nz/sites/default/files/Transient population report FINAL.pdf.

Katzenellenbogen, Judith M., Frank M. Sanfilippo, Michael St Hobbs, Matthew W. Knuiman, Dawn Bessarab, Angela Durey, and Sandra C. Thompson. 2013. "Voting with Their Feet - Predictors of Discharge against Medical Advice in Aboriginal and Non-Aboriginal Ischaemic Heart Disease Inpatients in Western Australia: An Analytic Study Using Data Linkage." BMC Health Services Research 13 (1): 1. https://doi.org/10.1186/1472-6963-13-330.

Katzenellenbogen, Judith M., Theo Vos, Peter Somerford, Stephen Begg, James B. Semmens, and James P. Codde. 2011. "Burden of Stroke in Indigenous Western Australians: A Study Using Data Linkage." Stroke 42 (6): 1515–21. https://doi.org/10.1161/STROKEAHA.110.601799.

Leather, Frances. 2009. "Prioritising Ethnic Data from Schools: Who Are We Missing? A Research Note." Social Policy Journal of New Zealand, no. 36: 82–87.

Marek, Lukas, Melanie Tomintz, and Malcolm Campbell. 2019. Population Transience and Healthcare Service Utilisation in Lakes District Health Board. Report for the Ministry of Health. University of Canterbury.

McCallum, Lisa K., Bette Liu, Peter McIntyre, and Louisa R. Jorm. 2014. "Estimating the Burden of Pertussis in Young Children on Hospitals and Emergency Departments: A Study Using Linked Routinely Collected Data." Epidemiology and Infection 142 (4): 695–705. https://doi.org/10.1017/S0950268813001039.

Mittelstadt, Brent Daniel, and Luciano Floridi. 2016. "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts." Science and Engineering Ethics 22 (2): 303–41. https://doi.org/10.1007/s11948-015-9652-2.

Raghupathi, Wullianallur, and Viju Raghupathi. 2014. "Big Data Analytics in Healthcare: Promise and Potential." Health Information Science and Systems 2 (1): 1–10. https://doi.org/10.1186/2047-2501-2-3.

Stats NZ. 2017. "Integrated Data Infrastructure: Overarching Privacy Impact Assessment." Accessed November 10, 2019. https://www.stats.govt.nz/assets/Uploads/Integrated-data-infrastructure/idi-overarching-pia.pdf.

Stats NZ. 2019a. "Research Using Stats NZ Microdata." Accessed November 10, 2019. https://cdm20045.contentdm.oclc.org/digital/collection/p20045coll17/search.

Stats NZ. 2019b. "Statistical Standard for Meshblock." Accessed November 10, 2019. http://archive.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/meshblock/definition.aspx

Stats NZ. 2019c. "Integrated Data Infrastructure." Accessed November 10, 2019. https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/.

Stewart, Robert, and Katrina Davis. 2016. "'Big Data' in Mental Health Research: Current Status and Emerging Possibilities." Social Psychiatry and Psychiatric Epidemiology 51 (8): 1055–72. https://doi.org/10.1007/s00127-016-1266-8.

Timmins, Kate A., Mark A. Green, Duncan Radley, Michelle A. Morris, and Jamie Pearce. 2018. "How Has Big Data Contributed to Obesity Research? A Review of the Literature." International Journal of Obesity 42 (12): 1951–62. https://doi.org/10.1038/s41366-018-0153-7.

Vogel, Christina, Stephen Zwolinsky, Claire Griffiths, Matthew Hobbs, Emily Henderson, and Emma Wilkins. 2019. "A Delphi Study to Build Consensus on the Definition and Use of Big Data in Obesity Research." International Journal of Obesity. https://doi.org/10.1038/s41366-018-0313-9.

Young, Angela, and Felicity Flack. 2018. "Recent Trends in the Use of Linked Data in Australia." Australian Health Review 42 (5): 584–90. https://doi.org/10.1071/AH18014.