Virtual Health Information Network recommendations to support the Health Research Council Assessing Committees

6 September 2018, Version 1.0

Sheree Gibb[1], Andrea Teng[1], James Stanley[1], Andrew Sporle[2], Barry Milne[2], Lucy Ashby[3], Gabrielle Davie[1], Nevil Pierse[1], Colin Simpson[4], Tony Blakely[1], Jeroen Douwes[5]

1. University of Otago
2. University of Auckland
3. Statistics New Zealand
4. University of Victoria Wellington
5. Massey University

Outline

Executive summary

This document is authored by expert members of Virtual Health Information Network (VHIN), and is intended to provide some background information about the Integrated Data Infrastructure (IDI) for HRC assessing committees. It describes the VHIN, the IDI, highlights the benefits of the IDI for health research, and identifies

some key elements that reviewers and/or assessing committees might want to know about the IDI when considering proposals that plan to use these data. Well-designed research in the IDI has the potential for high value health research output.

Committees assessing applications (either at full or Expression of Interest stage) that use IDI data will ideally include researchers experienced with the IDI. This document is intended to supplement (not replace) IDI expertise on the assessing committees.

What is the VHIN?

In 2015, with the expansion of health data in the StatsNZ IDI, the VHIN was established to create and sustain an environment that captures value from health data collections and other social and economic data. The VHIN is a network of over two hundred researchers, analysts and other professionals who use health and social data to generate insights that support the health and wellness of all New Zealanders. The network facilitates sharing and collaboration amongst network members in order to enhance health research outputs and improve health service delivery and health outcomes in New Zealand. Funding for the network has come from the Healthier Lives National Science Challenge, Health Promotion Agency, University of Auckland, Massey University and University of Otago. The VHIN is led by an Executive comprising University, Māori, Ministry of Health, and Statistics New Zealand stakeholders (VHIN Executive). Further information available at: https://vhin.co.nz/about/vhin-about/

What is the IDI?

The IDI is a large research database through which national health datasets are linked, using deterministic and probabilistic methods, to other government administrative data, survey data, census data and some NGO data. The IDI is managed by StatsNZ (formerly Statistics NZ) and protected under the 'five safes framework' where access is only allowed for trusted researchers and approved projects that are for the public good. Adding in data to the IDI is probably one of the highest priority areas to improve research quality in the IDI, especially if that data can be used for further research projects that follow. Further information on data available in the IDI is available here: https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Benefits of using IDI data

Over the last few years there has been increasing interest from health researchers in using IDI data. The linking of whole population data across different sectors opens up avenues of research that were difficult or impossible previously. Many of the benefits of using the IDI are similar to the benefits of research using linked administrative health data (like the National Minimum Data Set of hospital discharges) with the addition of datasets from multiple agencies. Some of the major advantages of IDI data for health research include:

- Linking information across sectors. Many national health datasets are now linked to national longitudinal data on education, social service use, justice, income, housing, census, border movements, and several StatsNZ surveys.
- Whole population data analysis. Many datasets in the IDI have national coverage and contain service use data for the whole population of New Zealand. The large sample sizes allow for analysis of small groups and rare events in ways that are not possible in projects that are dependent on primary collection of new data for that study (including surveys and other observational studies).
- Longitudinal data. Many of the datasets in IDI have a natural longitudinal structure, allowing researchers to follow cohorts of individuals over time.
- No recall bias. Administrative data eliminates the risk of recall bias, which is a problem when reliant on self-reports of service use (e.g. hospitalisation or pharmaceutical dispensing).
- Linking children and mothers. Using birth records in IDI it is possible to link (New Zealand born) children to their mothers. This provides information about a child's early environment (including during gestation).

- Security and confidentiality of the underlying data are carefully considered and are managed by an experienced agency (StatsNZ).

Things to look for in applications planning to use IDI data

Overall, the IDI is an important potential tool in health research and just like any other scientific tool knowledge and previous experience is the key factor for quality research. This section contains some suggested points to consider when assessing projects that plan to use IDI data.

Research Impact

There is a StatsNZ expectation that IDI users will share their code on the IDI Wiki (a Wiki-based platform for sharing information that is accessible to users with the StatsNZ Datalab system), and also externally to the Datalab (eg on the Virtual Health Information Network website). Transparency, code sharing and community building will support future IDI users in answering their research questions in a manner consistent with the public good requirements of StatsNZ data access.

Projects that will contribute to understanding the quality of IDI data are especially valuable. Information about data quality can be shared via IDI User forums, and through organisations like the Virtual Health Information Network.

It is a StatsNZ requirement that research from the IDI is published, and also made available to StatsNZ to share on their online research database. This is to maintain transparency and trust about how the IDI is being used, and so future researchers can see how they could use the IDI for their work.

Design and methods considerations

Reviewers should be confident that the researchers understand the following study design issues. Many of these concerns are common to studies that (a) use administrative datasets, like the Ministry of Health's national collections; or (b) are reliant on secondary analysis of other existing datasets (e.g. analysis of data collected for longitudinal studies like the Dunedin Multidisciplinary Study, or Growing Up in New Zealand).

- A description of the datasets and variables that the study plans to use: Some datasets have good metadata available; others not so. Discussion of the project with key dataset providers is recommended to identify strengths and weaknesses of the data researchers plan to use.
- What time periods are available for analysis: Some datasets have a short time series as data is not available across the entire IDI time-frame and this will reduce follow-up length and sample size.
- How the study population will be defined: There is an estimated resident population that may be useful for identifying the general population. Other projects may select a population based on who accessed a particular service.
- How comparison/denominator populations will be defined: This includes either appropriate identification of a comparison group (e.g. unexposed people in a cohort study type design) or the appropriate population denominator (e.g. all people with a particular [health condition] or the StatsNZ supplied estimated resident population).
- How much missing data is expected and how it will be managed: Not all variables are available for all people in IDI. Many variables are incomplete or have missing values due to missing data in the source records, or incomplete linkage between different agency datasets.
- Approach to linkage bias: False positive and false negative linkages may bias study results, and furthermore linkage rates may differ between study groups. It is important researchers understand how their study datasets are linked together, ie probabilistically via the IDI spine. Further information on the IDI spine and linkage rates can be found here; https://vhin.co.nz/guides/idi-spine/

- Most of the health data in the IDI is originates from health service use data. Consequently the IDI only contains information about health conditions/states/injuries where people have engaged with and been recorded by health services. Information about health states or events that have not engaged with services is missing from the IDI. As such, careful interpretation is needed of time trends, sub-group differences, and associations between variables. For example, if an increase in the use of mental health medications is found, it is unclear whether this is because more people have mental problems, or more people with problems are seeking treatment. Similarly, ethnic differences could be due to ethnic differences in prevalence, or service barriers between ethnic groups; and association between putative risk factors and disease/disorder (as defined by service use) may be biased by the factors influencing service use.

Practical issues with working in the IDI

- Researchers will need access to a secure Datalab. Researchers in Auckland, Wellington, Christchurch can access StatsNZ Datalabs. Researchers in other places may be able to use an institutional Datalab (for example, at various government agencies in Wellington, and at the University of Auckland, Auckland University of Technology, University of Waikato, University of Otago Wellington, University of Otago Dunedin, and Massey University, Wellington). Researchers should indicate how they are planning to access a Datalab. It is possible to set up a new Datalab but this is a long process and is subject to StatsNZ approval.

- If researchers are using specialist software (other than MS Office, SAS, Stata, R or SQL) they should indicate that they have discussed this requirement with Stats NZ and the software can be made available in the Datalab environment;

- Proposals using computationally intensive methods (e.g. machine learning) should include some reassurance that there will be sufficient computing power available in the Datalab environment to run these methods.

- Adding new data to the IDI is probably one of the highest priority areas to improve research quality in the IDI (e.g. adding primary care data to the IDI from a PHO; or social support provided by NGOs not otherwise already included in the IDI). This may be a lengthy process due to a range of factors needing to be considered and addressed, before the data goes through to a prioritization stage for loading into the Datalab. If researchers are planning to add new data to IDI, they should indicate that they have had initial discussions with Stats NZ about this and give some indication that approval is likely and how the timeframe fits with their research plans [1]. Final approval rests with Stats NZ.

- Realistic timeframes, particularly for new users: The IDI should be considered as a 'research tool' for which time is required to become familiar with. It may appear that having data already collected is time-efficient, however, it is important to realise that considerable time is required to understand potential 'trips and traps' of the administrative data in the IDI and to identify ways to mitigate these, where possible.

Some things are very difficult to measure with existing IDI data. This varies depending on the cohort and time period. Things that can be measured reliably at one point in time may not be available for other time periods. Similarly, measures that have poor coverage at population level may have good coverage for some population subgroups. Researchers should give some assurance that the variables that they want to measure are available for the population and time period that they are interested in. It is not possible to cover all the details about what is and is not possible with IDI data in this document- this is one area where experienced users can be useful to assessing committees.

Some common variables for health research are not readily available in IDI. BMI is only available for 4-year-olds, smoking is only available for people completing the 2013 census, and there is very little information about primary care service provision (e.g. visit information, diagnoses recorded in primary care records) as this is currently not part of the core IDI.

It is not reasonable to expect researchers to already have StatsNZ approval for IDI data access at the HRC application stage. Most projects will be approved for access so long as they contribute to the public good and meet StatsNZ's 'five safes' framework. It is expected that most projects being assessed by HRC would meet these criteria.

[1] Until July 2019, there are limited circumstances for adding new data to the IDI. Datasets will need to have a system-wide, high-priority, important and immediate policy need. This is not expected to affect projects in the current (2018) HRC round.

Responsiveness to Māori

Working with existing IDI data limits the potential of projects to be responsive to Māori because choices are limited to the existing datasets and variables they contain.

An important step is to consider the difference in coverage for Māori of the IDI datasets to be used. Some datasets have linkage rates that are lower for different ethnic groups. It is important that the extent and nature of missing data is evaluated, including when links cannot be found for a particular person. The impact of differences in missing data and data quality needs to be considered and tested with respect to how it might affect ethnicity related results. For example, missing tertiary education records might indicate either a failure to link across datasets or that a person had not participated in tertiary education and was thus not recorded in that dataset.

Consideration should be given to how ethnicity will be identified and researchers should apply the best practice methods for the data sources to be used. The IDI provides a source ranked ethnicity variable that prioritises the best quality data sources of ethnicity which may be useful. In all instances the approach for using ethnic identifiers should be described in a research application. Further information about using ethnic identifiers is available here; https://vhin.co.nz/guides/ethnicity-and-the-idi/

Information about Māori descent within the IDI is currently limited to the Census 2013 data set. The primary source of iwi identifiers is the Census 2013 dataset, with some information for younger age groups available from other data sources including education data. The iwi classification does not include all iwi and the classification was amended for the 2018 Census, resulting in some inconsistencies over time. Any intended use of iwi data needs to be planned in consideration of data governance principles associated with data about specific Māori groups (see www.temanararaunga.maori.nz for more information).

Expertise and track record

IDI expertise and experience is one of the most important factors in correctly using the IDI. The project should outline what support the IDI researchers will have from experienced IDI users and particularly individuals who are familiar with the datasets that will be used (either IDI users, or current/former staff members at particular Government agencies). Ideally the team should include an experienced IDI user as a named investigator or advisor (and include them on the Datalab application). Relying on existing training resources (the IDI Wiki, StatsNZ support) may not be enough for novice users, and this approach may run the risk of wasting time in getting started or making fundamentally incorrect decisions that can compromise the research outcome. Opportunities within a project for the development of the potential IDI workforce in terms of numbers or skills should be considered as resources and timeframes allow.

The research timeline should be carefully considered and may benefit from the involvement of experienced IDI users. The complexities of the IDI environment mean some tasks take longer than anticipated and experienced users are likely to be aware of resources and methods that may reduce workloads. The work plan should also include the time and resources required for any training and/or supervision of less experienced members of the research team.

Summary

We have described the VHIN, the IDI and highlighted the benefits of the IDI for health research, particularly though loading in additional data such as from primary care. We identify some of the issues that should be

managed when designing IDI research. Many of these considerations are relevant to research outside the IDI that uses administrative data or longitudinal cohort data. Well-designed research in the IDI has important potential for high value health research to inform improvements in policy and health services in New Zealand.

Further information

The Virtual Health Information Network is an excellent source of technical advice for IDI users, and provides short reports with technical information such as the best sources of ethnicity data, the linkage process and vital considerations in study design.

The StatsNZ Integrated Data webpage (including links to data safety considerations)

Advice on Ethics and the Integrated Data Infrastructure has been discussed by Dr Monique Jonas, confirming the importance of following institutional ethics processes for IDI research projects.

Te Mana Raraunga, for principals of data governance www.temanararaunga.maori.nz

The StatsNZ online research database has information about all integrated data research projects and publications.

 This work is licensed under a Creative Commons Attribution 4.0 International License.