



Options for meeting longitudinal information needs

A preliminary investigation



Crown copyright ©

This work is licensed under the [Creative Commons Attribution 3.0 New Zealand](#) licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to Statistics NZ and abide by the other licence terms. Please note you may not use any departmental or governmental emblem, logo, or coat of arms in any way that infringes any provision of the [Flags, Emblems, and Names Protection Act 1981](#). Use the wording 'Statistics New Zealand' in your attribution, not the Statistics NZ logo.

Liability

While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics New Zealand gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Citation

Statistics New Zealand (2014). *Options for meeting longitudinal information needs: A preliminary investigation*. Available from www.stats.govt.nz.

ISBN 978-0-478-40884-3 (online)

Published in March 2014 by

Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact

Statistics New Zealand Information Centre: info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4610
www.stats.govt.nz



Contents

List of tables	4
1 Executive summary	5
2 Introduction	6
Background.....	6
3 Longitudinal studies in New Zealand	8
Statistics NZ's longitudinal surveys and databases	8
Other departments' longitudinal dataset projects.....	9
Longitudinal data's contributions to research on New Zealand society	9
4 The four main types of longitudinal information need	10
Stability and change	10
Social transitions and gross flows	11
Cause and effect.....	11
Intergenerational mobility and transfers	12
Combined information needs.....	12
5 How well can existing data meet longitudinal information needs?	13
Linked administrative data.....	13
Linked census.....	15
Linked survey and administrative data	16
6 Options for collecting longitudinal data through surveys	20
Create or repeat a general-purpose panel survey	20
Introduce a longitudinal element into an existing cross-sectional survey	21
References	22
Appendix 1: Studies using Census Longitudinal Study in England and Wales	23
Demographics.....	23
Health.....	23
Housing.....	24
Education and social class	24
Work.....	24
Appendix 2: Summary of longitudinal information needs by domain, and rating of how well existing data can meet need	25



List of tables

- 1. Type of longitudinal data resource by its ability to inform longitudinal information needs 18
- 2. Summary of longitudinal information needs by domain, and rating of how well existing data can meet need..... 25



1 Executive summary

The aim of this paper is to identify the main future requirements for longitudinal data for research and policy purposes, and to discuss how well existing data sources could meet these requirements, as well as options for addressing the gaps. Our focus is on longitudinal information needs relating to individuals, families, and households. Information needs relating to businesses are outside the scope of this paper.

1. One of the major values of longitudinal datasets is providing information to describe life course trajectories and explaining variation in outcomes across the life course. As New Zealand society has become more dynamic and complex, some of the more predictable features of people's lives have given way to greater fluidity and diversity. This can be seen, for example, in people's working and family lives. As a result, cross-sectional data can no longer meet all the government's information requirements.

2. The review of Tier 1 statistics reflects this. The review identified income dynamics and worker turnover as important statistics. Statistical reviews of education, housing, families, and economic standard of living have identified a range of other relevant information needs requiring longitudinal data.

3. Four types of need for longitudinal data can be identified:

- information about stability and change in people's lives
- information about social transitions and flows, including life course transitions
- information to explain cause-and-effect relationships and interpret social change
- information on intergenerational mobility and transfers.

4. Longitudinal data are expensive to collect, and New Zealand has a small pool of researchers with experience in longitudinal data analysis. We need to avoid collecting more data than can actually be absorbed. We should explore reusing existing data sources before initiating new surveys.

5. Three main sources of existing data can potentially be used to inform longitudinal information needs:

- linked administrative data
- linked census
- linked survey and administrative data.

6. While each of these three sources has its strengths and limitations, together they could provide insights into each of the four identified areas of information need. They could inform, to a degree, these information needs across a wide range of subject-matter areas, including population mobility, family formation/dissolution, health, education, housing, criminal justice, work, income, and social mobility.

7. Despite the considerable benefit of reusing existing data, they will never address all longitudinal information requirements. Two survey options for meeting longitudinal information needs that can't be met through reuse of existing data are to:

- conduct a stand-alone longitudinal survey
- add a panel to an existing cross-sectional survey.

The feasibility of meeting information needs by adding a panel to an existing cross-sectional survey should be investigated, before initiating any further stand-alone surveys.



2 Introduction

Options for meeting longitudinal information needs: A preliminary investigation identifies future longitudinal data requirements in New Zealand and discusses which of these could be met by reuse of data from existing sources, and which might require a new survey. We will discuss methodological options for satisfying longitudinal needs that can't be met through reuse of existing data sources. Longitudinal data track the same individuals over multiple points in time, telling the story of how human development and changing times unfold.

Two key sources of longitudinal data are ending, so it is timely to review the demand for longitudinal information and consider options for meeting needs. These sources are the Longitudinal Survey of Immigrants: New Zealand, which completed fieldwork at the end of 2009, and the Survey of Family, Income, and Employment (SoFIE). Data collection for the final wave of SoFIE finished at the end of 2010.

It is becoming harder to obtain high response rates in official surveys. People are becoming busier, they do not understand the relevance of survey questions, and they fear government using their personal information. At the same time, government departments are under increasing pressure to reduce – or at least contain – costs.

It is therefore important we get maximum value from existing data sources before collecting new data. A range of administrative and survey data is available from government departments. We need to explore the potential for meeting longitudinal information needs from these sources, including through linking data across collections. Advances in computer technology and record linkage accuracy have significantly increased the ability to integrate data from multiple sources, and fill information gaps without increasing the burden on respondents.

This paper begins with background to the increased policy interest in longitudinal data, and describes existing sources of longitudinal data in New Zealand. We identify four main types of need for longitudinal information, provide examples of each type, and assess the potential for meeting them by reusing existing data. We then identify information needs that can't be met from existing data sources, and discuss survey options for meeting these.

Technical and policy issues associated with linking data, and issues of confidentiality and access are beyond the scope of this paper.

Background

In common with most western societies, New Zealand has become increasingly dynamic. For example, we have a geographically mobile population, greater volatility in working careers, and increasingly changeable family relationships. Whereas in earlier generations, most people's lives followed a fairly set path, today lives are less regimented and increasingly erratic. It is common for people to experience several partnerships over their lifetime, to change jobs multiple times during their working life, to move into and out of the labour force, to move between employment and unemployment, and to shift residence on a regular basis.

As a result, developments in society are less predictable. The statistical implication of a more dynamic society is that static descriptions have become less meaningful.

The shift to a more dynamic society has been accompanied by a shift to evidence-based policy, bringing with it a demand for a more robust foundation for understanding social change, and its causes and consequences. In addition to answering questions about

'how many?' or 'how much?', policymakers want to know 'what is going on?' and 'what works?' This requires information on the processes that are driving changes in society.

Aggregate statistics fall short of providing this information. While repeated cross-sectional data can measure broad changes at a macro level, they cannot help us understand change at the individual level. Time series of aggregate statistics may suggest relative stability, because most of the heterogeneity and dynamics remain hidden.

Longitudinal data track the same individuals over time. Not only does this enrich understanding about how society is changing, but it helps us understand why changes may or may not occur. Evidence from different points in time help us to unscramble cause and effect, and the connections between policies and outcomes. This can improve the development and evaluation of policies intended to prevent adverse outcomes or promote beneficial ones (Statistics New Zealand 2010c).

Longitudinal data are essential when:

- the phenomena of interest are directly concerned with individual-level change, that is, when time is part of the definition – for example, the dynamics of poverty
- investigating causal processes, including inferring cause and effect from temporal ordering or the effect of events
- an unmeasured characteristic (eg motivation or preferences) may affect the outcome, provided the unmeasured characteristic can be assumed to be fixed over the period of observation
- disentangling age, period, and cohort effects
- assessing the effect of policies on outcomes.



3 Longitudinal studies in New Zealand

New Zealand has a long association with longitudinal studies. The Dunedin Multidisciplinary Health and Development Study and the Christchurch Health and Development Study, both of which commenced in the 1970s, are recognised internationally as pioneering studies.

The studies, undertaken by the Dunedin and Christchurch schools of medicine, have each followed a birth cohort of children through childhood into adolescence and adulthood. The studies have been going for well over 30 years and data collection is continuing. They have resulted in rich databases of accumulated data. These databases are well suited to examining issues relating to changes in personal and family circumstances, and the linkages between these and a range of outcomes (for example, health, education, crime and substance abuse) in adolescence and young adulthood.

A major new survey of child development was launched in 2008. Growing Up in New Zealand (GUINZ) is a prospective pregnancy cohort study, which is based on a life-course epidemiological approach. It collects information on factors that influence child development through multiple contexts, and how these influences accumulate and change over time. The survey includes significant numbers of Māori, Pasifika, and Asian children, as well as European and other New Zealanders. Information from the survey will help us understand the complex factors that lead to differential child outcomes in an increasingly diverse population.

In addition to these cohort studies, academic institutions have carried out several other longitudinal studies in recent years. These include longitudinal studies of:

- Pacific peoples (Pacific Island Families: First Two Years of Life Study)
- Māori (Te Hoe Nuku Roa – Best Outcomes for Maori Study)
- ageing (New Zealand Longitudinal Study of Ageing).

Statistics NZ's longitudinal surveys and databases

Statistics NZ has conducted two major longitudinal surveys: The Survey of Family, Income, and Employment (SoFIE) and The Longitudinal Survey of Immigration: New Zealand (LisNZ).

The Survey of Family, Income, and Employment was designed to measure how New Zealanders' circumstances change over time and the factors that influence them. The eighth and final wave of data collection finished in 2010. The survey included alternating modules on net worth (assets and liabilities) and health status.

The Longitudinal Survey of Immigration: New Zealand provides information on migrants' settlement outcomes at 6, 18, and 36 months after taking up permanent residence in New Zealand. All three waves of data have been collected and output datasets are available for use.

Statistics NZ has also established longitudinal databases through linking administrative data. These include the Linked Employer Employee Database (LEED) and the Student Loans and Allowances Integrated Database.

LEED measures labour market and income dynamics. It was created by linking a longitudinal employer series from the Statistics NZ Business Frame to a longitudinal series of employer monthly schedule (EMS) payroll data from Inland Revenue. Subsequently, benefit data from the Ministry of Social Development (MSD) has been integrated into LEED.

The Student Loans and Allowances Integrated Database provides insight into the borrowing, qualifications, and income of students who have participated in the Student Loan Scheme or have received a student allowance. It was created by linking administrative data from the Ministry of Education, Inland Revenue, and StudyLink (Ministry of Social Development). The database is used to monitor the level of student debt and the speed with which debt is being repaid, and to cost and evaluate policies.

Two more examples of longitudinal databases Statistics NZ created by linking existing datasets are the Census Mortality Study and Cancer Trends. The Census Mortality Study links census and mortality records for the three years following the census to create cohorts for analysing the relationship between socio-economic factors and mortality. Similarly, the Cancer Trends study links census and cancer register records to provide information on the socio-economic risk factors for cancer.

Other departments' longitudinal dataset projects


Other departments, such as MSD and the Ministry of Health (MoH), have linked or are planning to link administrative data over time to create longitudinal files for research purposes. For example, MSD has assembled a longitudinal dataset from its benefit administration records. The Benefit Dynamics Dataset can show how long people spend on different types of benefit, and the prevalence and persistence of benefit dependence among New Zealand children.

MoH plans to link its administrative datasets so it can follow people's interactions with the health system from birth to death.

Longitudinal data's contributions to research on New Zealand society

These longitudinal studies and databases have resulted in an extensive body of research. This has expanded the knowledge base on the effects of social change in New Zealand, and factors influencing a range of social and economic outcomes among children and adults.

Research findings have informed policy developments in areas such as health, early childhood education, welfare, justice, and retirement income provision. For example, longitudinal research findings on savings informed the Review of Retirement Income Policy (Retirement Income Commission, 2010). Longitudinal research findings on welfare dependency have been used in the work of the Welfare Working Group (Welfare Working Group, 2010). The government established this group to make practical recommendations on improving the social and economic outcomes for people on a benefit.



4 The four main types of longitudinal information need

Statistics NZ has various ways of engaging with statistical users to determine their information needs. These include relationship meetings with government departments, regular advisory and user group forums, discussions relating to specific survey developments, and consultation in association with statistical reviews of subject-matter statistics (domain plans) and Tier 1 statistics.

We can identify four main types of need for longitudinal data from our engagement with users over recent years:

- descriptive information on stability and change in people's lives
- descriptive information on social transitions and flows
- information to explain cause-and-effect relationships
- information on intergenerational transfers and changes.

Stability and change

Information about stability and change at the individual level is needed in different life domains, such as health, employment, income, and housing. More specifically, information on how much time people spend in particular states, and the incidence of repeat occurrences of these states.

Cross-sectional information reveals how many people are in particular states at a given time. For example, it tells us how many people are unemployed, how many are disabled, and how many are in prison. However, it tells us nothing about the temporal nature of these states. In some instances, the states may be continuing, persistent, or even permanent, while in others they may be transitory or recurrent. Which of these circumstances is characteristic of the state can make a considerable difference, not only to the individuals involved, but also to the public policy significance.

Low income is an example. At the individual level, it matters whether the experience of low income is a brief and isolated episode, or whether it is long-term and/or recurs frequently over a person's life. At the societal level, it makes a difference to policy if low income is widely spread in the population, with most people having only a slight experience of it, or whether it is concentrated in a narrow group of individuals, most of whom have long and recurrent experiences of low income. The fact that income dynamics is a Tier 1 statistic reflects this topic's significance.

Some priority areas for information on stability and change include:

- income stability – what is the extent of income mobility among individuals and families? How sustained are transfers from low incomes to higher-income states? What are the characteristics of those individuals and families who experience an extended period in low income? What are the characteristics of those who cycle in and out of low income on a recurrent basis?
- welfare dependence – do the same individuals and families rely on government income support payments year after year, or do people move on and off welfare? Who is continuously reliant on government income support?
- employment instability – how often do people change jobs, and how does this vary by socio-demographic characteristics and across different occupations, industries, and regions?
- housing – how stable is housing tenure, and how does this vary by tenure type?

- partnership/family stability – is formal marriage more stable than de facto marriage? How many children spend time in a one-parent family during their childhood, and how long do they spend on average in a one-parent family?

Social transitions and gross flows

Much social change comes about from transitions between states, or changing rates of transitions between states. Often, too, the outcome is the result of two-way flows, with some people moving from X to Y, and others from Y to X.

Repeat cross-sectional data does not fully reveal these movements, and in fact conceals more than it illuminates. For example, cross-sectional data may show very little net change in the proportion of the population who are employed between two points in time, despite considerable movement of people into and out of employment. The outcome of these flows may be a significant change in the composition of the employed population, which may have implications for the most appropriate types of policy intervention.

Longitudinal data is needed whenever one is interested in identifying transitions and gross change at the micro level.

Some key types of transitions/gross flows that are of interest to researchers and policy-makers include those between:

- employment and unemployment
- government income support and work
- different family types (eg two-parent and one-parent family)
- partnered and unpartnered
- rental accommodation and home ownership
- different income groups.

There is also interest in the timing and sequencing of life transitions, including how they vary between different population groups, and how they are changing over time.

Particular life transitions of interest include:

- from education to work
- from employment to retirement
- from independent living to institutional care
- to parenthood
- to ownership of first home.

Cause and effect

Government is interested in the causes and effects of transitions, events, and outcomes. They need to understand the determinants of outcomes of policy interest (eg low income, poor health, unemployment, criminal offending) so they can identify the key points for intervention.

Cross-sectional data can identify whether an association exists between two variables, such as between unemployment and poor health, but it can't tell us which is caused by which. To confidently demonstrate the direction of causality between variables, we need longitudinal data. Because they follow the same individuals over different points in time, and the data have a time order of measurement, we can determine the causal association when the cause precedes the effect (or the outcome of policy interest).

However, establishing causality is not always straightforward. Some cause-and-effect relationships can take a long time to manifest themselves. The elapsed time can vary from a few months to decades. The connections can also be complex and cumulative (eg disadvantaged circumstances in early childhood leading to poor school performance, and eventually to poor labour market experience). To understand these connections, we need information on the trajectories of individual life histories.

Some identified areas of interest in terms of understanding cause and effect are:

- causal factors associated with variation in immigrant settlement outcomes
- causes and consequences of housing and geographic mobility
- links between educational qualifications and employment outcomes
- the effects of childhood circumstances on later life
- determinants and consequences of health outcomes, such as heart disease, cancer, diabetes
- factors that trigger a transition into or out of the low-income category
- impact of economic shocks on the well-being of New Zealanders
- effects of student loans on social and economic outcomes and life choices.

As well as determining the most appropriate points for policy intervention, government needs to understand the effects of its policies. This includes understanding what works, and what does not, as well as why, for whom, and in what contexts. In practice, it means measuring the difference between the extent to which outcomes that are the object of a policy are achieved when the policy is in place, and the extent to which they would have been achieved in its absence. For example, there is interest in evaluating the effect of KiwiSaver on savings and wealth accumulation.

Intergenerational mobility and transfers


The nature and extent of intergenerational mobility interests social scientists and policymakers because of its implications for equality of opportunity. Intergenerational mobility is the relationship between the circumstances of parents and the circumstances of their children. Low levels of intergenerational mobility can imply that people's backgrounds mean they are unable to fully develop and use their skills. Barriers to people developing and fully using their abilities could hinder skills development and productivity growth, and the achievement of improved living standards.

Much of the attention in relation to intergenerational mobility is on educational and occupational attainment. A particular area of interest is intergenerational mobility between first- and second-generation migrants. There is also growing interest in the persistence of welfare dependency and poverty across generations, the intergenerational transmission of family violence and criminal behaviours, and intergenerational cycles of teenage parenting. The prevalence of intergenerational transfers of wealth, and their importance in wealth accumulation and inequalities in wealth distribution, are also of interest.

Combined information needs

While we can distinguish between these four types of longitudinal data need, in practice some users require different combinations of them. The most common combination involves looking at cause and effect in conjunction with one of the other types of information need. Some examples include identifying the factors that:

- influence persistent or recurrent poverty
- result in families moving into or out of low income
- lead to intergenerational poverty.



5 How well can existing data meet longitudinal information needs?

This section discusses the extent to which reuse of existing data can meet the four main types of information need. The three main sources of existing data are:

- linked administrative data
- linked census
- linked survey and administrative data.

Linked administrative data

Many administrative datasets accumulate and store information about individuals over time, such as their taxable income, welfare payments, hospital episodes, educational attendance, and court convictions. Often they contain information that isn't well measured by survey/census data, such as program participation or health status, which is very policy relevant.

Administrative datasets are inherently longitudinal, as the information for each individual can be linked together to show change over time. Linked administrative records can provide important insights, particularly in sectors where there are many parts to the system, such as education, health, and justice. Linking across the system using unique identifiers such as the National Health Index (NHI) number in the health system, or Person Record Number (PRN) in the justice sector can result in a person-oriented database – that is, a database about individuals and their attributes or characteristics. For example, linking data on contact with the police, courts, and prison system permits analysis of issues, such as patterns of recidivism.

A key strength of these linked administrative databases is that they cover everybody who has had contact with the relevant government system or service, including people in institutions and other types of non-private dwellings. This means it is possible to analyse and dissect the data, without encountering sample error problems.

In addition, the data give a continuous account of changes in status, allowing precise measurement of time spent in a single spell and the incidence of multiple spells. As a result, the data are well suited to studying stability and change at the individual level in areas such as earnings, employment, benefit receipt, health (hospital care), and reoffending/recidivism. By comparison, longitudinal survey data are generally constrained to retrospectively recording status in monthly windows, or taking snapshots at annual intervals. Furthermore, survey data rely on respondents' recollection of events, so are subject to recall error.

Linked administrative data have some limitations. First, any one administrative database is representative only of the specific population related to the administrative purpose. Consequently, the data are in general not representative of the usually resident population. For example, the Benefit Dynamics Database only covers benefit claimants, so it provides only a partial view of children and families in low-income situations.

Second, some individuals and groups may avoid contact with government agencies, and so may be under-represented on administrative databases. For example, they may not claim benefits they are entitled to, or they may conceal their earnings. This can result in significant biases.

Third, administrative databases contain little information about the characteristics and circumstances of individuals. This limits their value for exploring causal influences and the relationship between outcomes and policy levers.

Fourth, information that is a by-product of administrative processes may not link neatly with our conceptual frameworks and statistical definitions.

Fifth, because administrative data are usually collected for purposes other than producing official statistics or research, the potential application and questions each dataset can address are usually constrained to one domain.

Administrative databases can be made more useful by linking records between sectors. The Linked Employer-Employee Database (LEED) is an example of such a database. LEED was originally created by linking data from the taxation system with business data from Statistics NZ's Business Frame. Subsequently, administrative data on benefits, and tertiary enrolments and completions (EOTE) have been added.

There are advantages to developing a more systematic approach to linking data within the Official Statistics System. Statistics NZ has developed a new flexible data integration environment (the IDI). With the IDI, we can create links between administrative datasets at the person level more easily, and manage them in a way that respects privacy and security protocols. The IDI environment enables existing administrative data to be used and reused for producing official statistics and applied to policy and research activities.

There is also value in linking sample surveys into this environment. The strengths of the linked administrative data lie in the ongoing accumulation of detailed information, well beyond the time period of any survey dataset and for less cost, permitting the identification of long-term trends and causal relationships, at disaggregated levels that allow subgroup analyses not possible with sample survey data.

LEED has demonstrated its value as an important source for statistics and policy-relevant research on labour market dynamics. The establishment of IDI allows more datasets to be integrated, allowing researchers to investigate a wider range of questions. Examples of datasets that could be integrated in the IDI environment include administrative data on:

- births and deaths
- health system
- early childhood education
- vulnerable children
- police apprehensions
- tenancy bonds
- citizenship.

The linked data would provide the flexibility to respond rapidly to new information needs. The data would be most useful for questions about a population that could be readily identified in the data. For example, beneficiaries, students, or migrants could be followed over time and across sectors. Some examples of potential uses include:

- tracking outcomes across sectors for children living in benefit-dependent households, eg health, education, employment, migration, and justice outcomes
- assessing the effectiveness of employment and education programmes for different types of beneficiaries, and the relationship with health outcomes
- identifying migrants' pathways through and outcomes in education, employment, benefit receipt, and health from the time they arrive in New Zealand
- measuring the effectiveness of policies designed to retain knowledge and skills in New Zealand, eg locally trained doctors
- evaluating the post-prison outcomes of offenders who participated in various programmes while in custodial care.

It is important to recognise that the IDI will not address all longitudinal information needs. In particular, it will not provide insight into the influences and effects of changes in family and household circumstances, such as a change in family income. It will not help us understand the determinants of household savings and wealth accumulation and the vulnerability of different types of household to external shocks. Intergenerational mobility is another information need it will not satisfy.

Other limitations of linked administrative datasets are that the data are not representative of the New Zealand population, the data are sometimes incomplete (eg qualifications only cover those gained in New Zealand), and linkage errors may affect the quality of the data and their use in research and analysis¹.

Linked census

Creating a linked census is another way to meet longitudinal information needs. A linked census is constructed by linking census records of individuals at successive censuses to create a longitudinal database. Often a sample of census records is linked to manage dataset size and allay confidentiality concerns. Administrative data on births, deaths, and migration are commonly added to the linked census, to adjust for additions to and departures from the population, and make sure the database represents the usually resident population. Other administrative and survey data may also be linked to increase the value of the database to users.

The Census Longitudinal Study (LS) in England and Wales is an example of a linked census. The LS contains linked census and vital events for 1 percent of the population of England and Wales. Information from the 1971, 1981, 1991, and 2001 Censuses has been linked together, along with information on events such as births, deaths, and cancer registrations. The entire census records for the LS member and all members of that person's household are entered into the LS. The original sample was selected in 1971 from four birthdates across the calendar year. New members with the same four dates of birth enter the study at birth, or when they immigrate to England and Wales. Members leave the study when they die or emigrate (Office for National Statistics, 2010).

The LS has supported diverse analyses, including demographics and health-related research, for which the database was originally set up. The major topics it has informed include:

- inequalities in health, employment, and education
- social mobility
- geographical mobility
- labour market mobility
- interethnic mobility
- family and partnership stability
- social exclusion
- movement between housing tenures
- migrant experiences
- youth transitions
- ageing
- fertility (childbearing patterns).

¹ The lack of a common identifier makes linkage across collections difficult. As a result, most links will be probabilistic.

The large sample size of the LS, which includes around 500,000 people at any one census, and the length of follow-up, have also made it a valuable source for statistically robust studies of population subgroups, such as selected ethnic and occupational groups. For example, it has been used to estimate the mortality rates of particular occupations and to study the socio-economic differences in the incidence of stomach, colorectal, and pancreatic cancers. Appendix 1 includes a list of studies that have used the LS.

The Australian Bureau of Statistics (ABS) has also established a linked census modelled on the UK linked census. They created the Australian Census Longitudinal Dataset (ACLD) by combining data from a 5 percent random sample of the population at the 2006 census with data from corresponding records at the 2011 Census. The ACLD will also be linked to several non-ABS datasets. These include the birth and death register (including cause of death), data on long-term migrants, and national disease registers (Australian Bureau of Statistics, 2010).

The establishment of a linked census in New Zealand will allow researchers to address many relevant policy-related questions that can't be readily answered with data from other sources. A linked census will facilitate studies that relate demographic, social, and economic outcomes to previous circumstances, including family and household circumstances. Thus, it will be possible to answer questions about what led to people being in a particular situation at a particular time, and the likely consequences of it on the rest of the person's life. Examples include:

- the effects of living in a deprived neighbourhood on education and employment outcomes in adulthood
- the influence of socio-economic status on childbearing patterns
- the cumulative effects of economic and social events and outcomes over the life course on ageing outcomes.

A linked census will also make it possible to study intergenerational mobility in New Zealand, assuming data can be linked across at least three censuses. It will provide information on subgroups of policy interest, and insight into how regional effects interact with patterns in individuals' experiences over time.

A linked census could be enriched by combining it with routinely collected administrative data. This would enhance the value of the linked census, without the extra burden of data collection. Proof of concept has been established by linking death records and cancer registration data to the census, using probabilistic matching². Other administrative data that could potentially be linked using probabilistic matching include education, welfare and tax records, new migrant approvals, and ACC claims.

A linked census will have less value if New Zealand moves to a 10-yearly census cycle from the current five-yearly cycle. Nevertheless, the experience of England and Wales – which have a 10-yearly census – indicates it is still worth doing.

Linked survey and administrative data

Another option for creating a longitudinal data resource from existing data is by linking cross-sectional survey data to administrative data. By linking these sources, we can take advantage of the strengths of each.

Surveys provide a nationally representative sample and detailed background information on the demographic, social, and household/family characteristics of participants and non-participants of government programmes and services. Administrative sources provide

² Matching was done using information available on both the census and death records, including date of birth, sex, country of birth, ethnicity and area of usual residence.

reasonably reliable data on individuals' participation in government programmes and services, potentially for long periods.

Linking survey and administrative data combines the range and scope of the survey results with the completeness and accuracy of the administrative records. The matched data allows researchers to study the relationship between survey responses and administrative programme/service participation before and after the survey, thereby providing a longitudinal component to the cross-sectional survey.

Statistics NZ has linked data from the Household Labour Force Survey (HLFS) with administrative data on earnings in the IDI. This allows researchers to study survey respondents' work and earnings dynamics before and after their interviews, and investigate how these dynamics relate to explanatory factors, such as ethnicity and family circumstances, available from the HLFS. The IDI also contains data from the New Zealand Income Survey; the Survey of Family, Income, and Employment; and the Longitudinal Immigration Survey.

There is scope for linking other surveys to administrative data to create longitudinal data resources. For example, the New Zealand Health Survey could be linked to administrative information about the use of health services and to disease registers and death records. Researchers could then investigate the relationship between health status and using health services, and give insight into differences in how people use these services. It would also make it possible to examine whether self-reported measures of health behaviour and health status can accurately estimate the risk of disease and death in the New Zealand population.

Similarly, the Household Economic Survey could be linked to administrative data on tax and income support, to provide detailed data on income and welfare dynamics of the survey respondents. The linked data would improve the accuracy of predictive modelling systems that help policymakers understand the broad effects of policy changes.

There are substantial analytical benefits to linking administrative and survey data. One major advantage is being able to compare programme participants with non-participants, which is not possible when analysing administrative data on their own. Another advantage is the wide range of explanatory variables that are available from the survey data to investigate the correlates of social outcomes, such as the correlates of individuals and families who have low income over an extended period.

Linked survey and administrative data, however, cannot inform all types of longitudinal research questions. In particular, the data are unsuitable for studying the causes and consequences of transitions between states, for example into or out of low income, or between different family types. They are also not suitable for understanding intergenerational transfers and the impact of accumulated life experience on later outcomes, such as the impact of childhood experiences on outcomes in youth and adulthood.

Table 1 summarises the data options discussed above and their usefulness for informing the four types of longitudinal information needs. Appendix 2 summarises information needs by domain, and indicates which of these needs could be met by existing data.

Table 1
Type of longitudinal data resource by its ability to inform longitudinal information needs

	Linked administrative data	Linked census and administrative data	Linked survey and administrative data
Stability and change	✓✓✓	✓✓	✓✓✓
Social transitions and flows	✓✓✓	✓✓	✓✓✓
Understanding cause and effect	✓✓	✓✓✓✓	✓✓✓
Intergenerational mobility and transfers		✓✓✓✓	

The main findings are:

- Existing data sources could address a wide range of longitudinal information needs.
- Reuse of existing data could inform questions about the temporal nature of states (ie whether they are continuing, temporary, or transitory). However, the usefulness of linked administrative data for this purpose could be affected by coverage issues, which may result in significant biases in the data. Linking survey with administrative data could overcome this problem, assuming that the linking process does not cause biases. The value of linked census data for understanding the temporal nature of states would be limited because it would not provide a continuous record of change, although adding administrative data could help address this.
- Each of the three data options could provide information on transitions and flows, although again the usefulness of the linked census would be limited because it would only provide information at census dates.
- Reuse of existing data could provide some insights into causality. A linked census would contain a range of information on the demographic, social, and family/household characteristics of individuals over a long period. This could show the factors that affect outcomes in the shorter and longer term. However, the linked census would be less useful for providing insight into the factors that trigger transitions, such as the factors causing a transition to low income. This would also be a limitation of linked survey and administrative data. The usefulness of linked administrative data is limited by the relatively small number of explanatory variables

that is available, and by the fact that it is often not representative of the usually resident population.

- The linked census would be the most suitable option for studying intergenerational transfers, although it would be dependent on the ability to effectively link data over at least three censuses.
- A limitation of all three options is that they are not well suited to studying family/household transitions, and their related causes and effects. The linked census would provide information on transitions between census dates, but would miss those that occur in the intervening period. Neither linked administrative data nor linked survey and administrative data would provide information on family/household transitions.
- All three options are limited by the range of variables available from existing sources, so cannot address some topics of policy interest, such as the factors that influence savings behaviour and the accumulation of net worth. Also, the existing sources may not capture information on important explanatory factors for some outcomes.

The next section considers survey options for addressing these gaps and limitations.



6 Options for collecting longitudinal data through surveys

There are various methodologies for collecting longitudinal data in surveys. These include retrospective surveys, cohort studies, and representative panel surveys.

Retrospective surveys interview respondents once about events in their past. Because respondents do not easily recall past events, retrospective surveys are only appropriate for measuring significant and infrequent life events, such as births and marriages. People are unlikely to accurately remember earnings or income levels beyond the immediate past.

Cohort studies typically focus on people comprising a specific birth cohort of the population. A birth cohort is at any time representative only of one age group, whose experiences are influenced by the historical times at which they reach critical transitions. This limits their value for many purposes, including study of the impact of policy changes, unless there are other not-too-distant cohorts available for comparison.

Because of the limitations of retrospective studies and cohort studies, we will not discuss them further in this paper. Instead, we present two options we consider more appropriate for informing unmet longitudinal information needs. They are:

- a general-purpose panel survey
- introducing a rotating panel design to an existing cross-sectional survey.

Create or repeat a general-purpose panel survey

Large continuing longitudinal surveys, with wide-ranging subject matter, are a strategic resource for policy-relevant research. They tend to focus on the main, long-established topics (such as employment, income, family, and health) which are of common importance to many areas of policy and have a cross-disciplinary interest. These types of survey are common internationally. Examples include the Panel Survey of Income Dynamics in the USA, Survey of Labour and Income Dynamics in Canada (SLID), British Household Panel Survey in the UK, European Union Survey on Income and Living Conditions (EUSILC), and the Household Income and Labour Dynamics in Australia Survey (HILDA).

In New Zealand, the Survey of Family, Income, and Employment (SoFIE) is an example of a general-purpose panel survey. It collected data on some 22,000 individuals, across hundreds of survey items of information, for eight years. The survey covered the main topics of employment, income, family, health, assets, and liabilities. The waves 1 to 8 dataset was only recently released for analysis, so it is too early to assess the value of the survey data for policy development and evaluation. Some examples of research topics the data are being used to inform are:

- impact of labour market factors, asset wealth, income, and family dynamics on health
- impact of home ownership on labour market outcomes, health status, wealth, educational status, and transitions to employment for youth
- evaluating KiwiSaver's effect on New Zealanders' savings behaviour
- adequacy of households' savings and wealth accumulation for retirement, the principal determinants of net worth, and households' vulnerability to economic shocks that would affect their ability to service debt

- impact of economic shocks on New Zealanders' well-being, including the characteristics of individuals prone to economic shocks, the impact of these shocks on employment status, wages, and income, as well as on several other individual outcomes.

A repeat of SoFIE would satisfy most of the key needs of users that reusing existing data could not meet. This includes longitudinal data on families' net worth, and information on the immediate causes and effects of significant transitions, such as the transition to low income. However, longitudinal surveys are very expensive to conduct, impose a heavy burden on respondents, and are complex to analyse. There are not enough researchers in academia and government in New Zealand with experience in analysing longitudinal data to fully exploit current volumes of longitudinal data. We need to avoid collecting more data than can be used. We also need to avoid collecting data that are too complicated for New Zealand researchers to analyse.

Before embarking on another major longitudinal survey, we must quantify the benefits of collecting new data, over and above the data that can be obtained from existing sources. We also need to explore reducing survey costs by collecting the data by phone or Internet, rather than interviewing in person, and using administrative data sources in place of direct collection. This will enable a decision to be reached on whether a future longitudinal survey is affordable.

Introduce a longitudinal element into an existing cross-sectional survey

An alternative option for meeting needs that existing data can't is to introduce a longitudinal element into an existing cross-sectional survey, such as the General Social Survey or the Household Economic Survey (Income). There are different methodological options for achieving this, including a rotating panel design and a split panel design.

A rotating panel is so called because individual panel members are rotated into and out of the panel over time. This design therefore involves a succession of separate probability samples of the population at different points in time. Each panel has a fixed life, although the survey itself may be indefinite.

Rotating panels can be used to measure the frequency and timing of events occurring in a given time period. However, the relatively short length of the panel limits how well rotating panels can illuminate long periods in low income or multiple transitions between states. The usefulness of the rotating panel design for analysing the effects of policies may be limited if the effects accumulate over long periods, or there are substantial lags. For policies with a more immediate impact, the effects should be measurable. The Household Labour Force Survey (HLFS) has a rotating panel design, but has not been used extensively for longitudinal analysis.

A split-panel design is a combination of a panel and a repeated cross-sectional component. The panel component can be used to measure individual change, including the frequency and timing of events, and transitions between states. Both components can be used to provide cross-sectional estimates. The British Social Attitudes Survey is an example of a split-panel survey (GJ Duncan & G Kalton, 1987).

Because some priority longitudinal information needs are about understanding family/household income dynamics and the factors affecting family/household net worth, changing the design of the Household Economic Survey (Income) to meet both longitudinal and cross-sectional needs would be more appropriate than adding a longitudinal component to the NZGSS. This option would be more cost effective than conducting another stand-alone panel survey.

Further work is needed to fully investigate the technical feasibility, and costs and benefits, of adding a longitudinal component to the Household Economic Survey (Income).



References

Duncan, GJ, and Kalton, G (1987). Issues of design and analysis of surveys across time. *International Statistical Review* 1987(55).

Office for National Statistics (2010). [The ONS Longitudinal Study of England and Wales](#). Available from www.ons.gov.uk

Retirement Commission (2010). [2010 review of retirement income policy](#). Available from www.cflri.org.nz

Statistics New Zealand (2007). [Report of the review of family statistics](#). Available from www.stats.govt.nz

Statistics New Zealand (2009a). [Review of crime and criminal justice statistics report 2009](#). Available from www.stats.govt.nz

Statistics New Zealand (2009b). [Review of housing statistics report 2009](#). Available from www.stats.govt.nz

Statistics New Zealand (2010a). *Review of economic standard of living report*. Unpublished report.

Statistics New Zealand (2010b). *Review of population statistics*. Unpublished report.

Statistics New Zealand (2010c). [Social and population statistics information needs for New Zealand: Towards 2020](#). Available from www.stats.govt.nz

Welfare Working Group (2010). [Long-term benefit dependency: The issues \(detailed paper\)](#). Available at www.ips.ac.nz



Appendix 1: Studies using Census Longitudinal Study in England and Wales

The LS has been widely used in a variety of research areas, most notably including demographics or health-based research for which the study was originally established. The main research areas include demographics, health, housing, education and social class, and work.

Demographics

The database has been used to study patterns of mortality and fertility in the UK in relation to other factors, such as social class. Studies have covered mortality, including cause of death, longevity and ageing, and fertility (including intergenerational patterns of teenage fertility). It has also been used to study events around the household and its composition. Examples of specific areas of study include:

- household circumstance and transitions
- stability and instability in children's lives
- changing partners
- transitions to institutional residences among older people
- housing demand and need generated by divorce
- migration during early family life
- gender, migration, and household change among older people
- risk factors for low birth weight
- socio-economic inequalities in cause-specific mortality.

Health

The database contains a range of health indicators from administrative data including infant mortality, mortality, cancer registration, and birth weight. In addition, there are questions in the census on self-reported health and long-standing illness. The database has spawned much research into differential mortality and health inequality. It has also been used to investigate geographical differences in health, occupational health, infant mortality, and cancer incidence and survival. Examples of specific areas of study include:

- relationship between social mobility and health outcomes
- teenage mothers and the health of their children
- persistence of health inequalities in old age
- risk factors and low birth weight
- migration and the health of older people
- temporal, social, and spatial effects of health inequality
- fertility history and health in later life
- relationship between unemployment and chronic illness.

Housing

The database has been used to investigate a variety of research topics related to housing. These include:

- the movement between housing tenures
- the relationship between residential migration and housing tenure
- housing demand and the need generated by divorces
- differences in mortality by housing tenure and car access
- labour and housing market change
- intergenerational change in the housing market.

Education and social class

A number of studies have used the database to look at the effects of education and social class on other outcomes, such as mortality and childbearing patterns. The database has also supported research on intergenerational mobility. Examples of specific areas of study include:

- association between parental social class and own social class
- educational inequalities amongst second-generation migrants
- social mobility of immigrants
- geographical mobility and social mobility
- social distribution of cancer
- role of family background and education in shaping social class outcomes among ethnic groups
- impact of gentrification on displacement of certain groups.

Work

The database has been used for research on a range of labour market outcomes. Examples include:

- accumulated labour market disadvantage and its impact on long-term illness
- socio-demographic origins and consequences of unemployment
- unemployment and mortality
- unemployment and fertility
- occupational sex segregation and part-time work.

Appendix 2: Summary of longitudinal information needs by domain, and rating of how well existing data can meet need

Table 2
Summary of longitudinal information needs by domain, and rating of how well existing data can meet need

Domain and topic	Information need	Linked admin data	Linked census and admin data	Linked survey and admin data	Birth cohort studies
Population					
Fertility	<ul style="list-style-type: none"> • number, timing, and spacing of births over life course • factors influencing the number, timing, and spacing of births 	✓	✓✓✓	✓	
Geographical mobility	<ul style="list-style-type: none"> • nature and extent of geographical mobility over life course • antecedents and consequences of geographical mobility 	✓	✓✓✓		✓✓
Migrant settlement	<ul style="list-style-type: none"> • settlement outcomes of migrants • factors that facilitate and hinder successful outcomes • outcomes of children of migrants 	✓✓✓	✓✓✓✓	✓✓✓	✓
		✓	✓✓	✓	✓
			✓✓		✓
		✓✓✓	✓✓	✓✓	✓✓✓
		✓	✓✓✓	✓✓✓	✓✓✓
Health					
Lifetime health	<ul style="list-style-type: none"> • health transitions and pathways over life course 	✓✓	✓✓	✓✓	✓✓✓
Health inequalities	<ul style="list-style-type: none"> • causes and correlates of health transitions and outcomes 	✓	✓✓✓	✓✓✓	✓✓✓

	<ul style="list-style-type: none"> • impact of health status on health utilisation patterns • effect of health interventions on health outcomes 			✓✓✓	✓✓
Knowledge and skills					
Social and economic returns on education	<ul style="list-style-type: none"> • learning trajectories and transitions over life course • determinants of learning participation and outcomes • effect of student loans on economic and social outcomes • intergenerational transfers of human capital 	✓✓✓	✓✓✓	✓✓✓	✓✓✓
		✓	✓✓	✓	✓✓✓
		✓	✓✓	✓	
			✓✓		✓✓
Paid work					
Labour force mobility	<ul style="list-style-type: none"> • extent of mobility into and out of the labour force • antecedents and consequences of labour market transitions • persistence of unemployment • intra- and intergenerational occupational mobility • earnings mobility 	✓✓✓	✓✓✓	✓✓✓	✓✓
		✓✓	✓✓✓	✓✓	✓✓
		✓✓✓	✓✓	✓✓✓	✓✓
			✓✓✓	✓	✓
		✓✓✓✓	✓✓	✓✓✓	✓✓
Economic standard of living					
Income and wealth mobility	<ul style="list-style-type: none"> • nature and extent of income mobility for individuals and families • persistence of low income and recurrent low income for individuals and families • persistence of welfare dependence and recurrent welfare dependence 	✓✓✓	✓✓✓✓	✓✓✓✓	✓✓
		✓✓✓	✓✓✓✓	✓✓✓✓	✓✓
		✓✓✓	✓✓✓	✓✓✓	✓✓

	<ul style="list-style-type: none"> • determinants of individual and family income transitions • nature and extent of wealth mobility • adequacy of households' savings and wealth accumulation for retirement • determinants of savings and net worth • intergenerational transfers of net worth 	✓	✓✓	✓✓	✓✓
Housing					
Housing careers	<ul style="list-style-type: none"> • dwellings and tenure mobility • antecedents and consequences of housing transitions • impact of home ownership on outcomes in other domains 		✓✓		✓✓
			✓		✓✓
			✓✓✓		✓✓
Safety and security					
Offender life history	<ul style="list-style-type: none"> • offending trajectories over lifetime • determinants of criminal offending • extent of recidivism • effect of criminal justice interventions on outcomes 	✓✓✓	✓✓✓		✓✓
		✓	✓✓		✓✓
		✓✓✓	✓✓✓		✓✓
		✓✓✓	✓✓✓		✓✓
Culture and identity					
Inter-ethnic mobility	<ul style="list-style-type: none"> • extent of mobility between ethnic groups • determinants of inter-ethnic mobility • impact of inter-ethnic mobility on growth of ethnic groups 	✓✓	✓✓✓✓	✓✓	✓✓
			✓✓✓	✓	✓✓
			✓✓✓✓		

Cross-cutting topics					
Family dynamics	<ul style="list-style-type: none"> • number and types of family transitions 		✓ ✓		✓ ✓ ✓
Social mobility	<ul style="list-style-type: none"> • duration of time spent in different family types 		✓		✓ ✓ ✓
	<ul style="list-style-type: none"> • antecedents and consequences of family transitions 		✓		✓ ✓ ✓
Social exclusion	<ul style="list-style-type: none"> • extent of intergenerational social mobility 		✓ ✓ ✓		✓ ✓
	<ul style="list-style-type: none"> • extent of intragenerational social mobility 		✓ ✓ ✓		✓ ✓
Child development	<ul style="list-style-type: none"> • extent of persistent multiple disadvantage 		✓ ✓		✓ ✓
	<ul style="list-style-type: none"> • causes and correlates of persistent multiple disadvantage 		✓ ✓		✓ ✓
Ageing	<ul style="list-style-type: none"> • impact of childhood experiences on child, adolescent, and adult outcomes 		✓ ✓ ✓		✓ ✓ ✓
Economic shocks	<ul style="list-style-type: none"> • combined effect of social, economic, and environmental factors and transitions on the well-being of individuals as they age 	✓	✓ ✓ ✓	✓ ✓	
Economic shocks	<ul style="list-style-type: none"> • impact of economic shocks on the outcomes of individuals and families 	✓	✓ ✓	✓	✓ ✓