Australian Government

Department of Social Services

Australian Institute of Family Studies

Growing Up in Australia:
The Longitudinal Study of Australian Children (LSAC)
LSAC Technical Paper No. 28

Growing Up in Australia

The Longitudinal Study of Australian Children

Combining linked data from NAPLAN and Medicare with LSAC survey data

Jennifer Prattley and Karlee O'Donnell

June 2023

Acknowledgements

This report makes use of data from *Growing Up in Australia*: The Longitudinal Study of Australian Children (LSAC). LSAC is conducted in a partnership between the Department of Social Services (DSS), the Australian Institute of Family Studies (AIFS) and fieldwork providers (the Australian Bureau of Statistics (ABS), up to and including Wave 9, and Roy Morgan, from Wave 10 onwards), with advice provided by a consortium of leading researchers.

The authors wish to thank Lisa Mundy, Amanda Vittiglia and the wider AIFS LSAC team for their contributions to the paper and project. The authors would especially like to thank LSAC participants for their ongoing contribution to the study.

The Australian Institute of Family Studies is committed to the creation and dissemination of research-based information on family functioning and wellbeing. Views expressed in its publications are those of individual authors and may not reflect those of the Australian Institute of Family Studies.

Growing Up in Australia: The Longitudinal Study of Australian Children is conducted in partnership between the Australian Government Department of Social Services, the Australian Institute of Family Studies (AIFS) and the Australian Bureau of Statistics (ABS), with advice provided by a consortium of leading researchers from research institutions and universities throughout Australia growingupinaustralia.gov.au.

Technical paper

The Longitudinal Study of Australian Children: LSAC Technical paper No. 28. Combining linked data from NAPLAN and Medicare with LSAC survey data.

Authors: Jennifer Prattley and Karlee O'Donnell

For more information, write to:
Longitudinal Studies – Research and Methods
Data Strategy and Development| Data and Evaluation
Australian Government Department of Social Services
PO Box 7576
Canberra Business Centre ACT 2610

Email: LongitudinalStudiesDataAccess@dss.gov.au

ISBN 978-1-76016-281-8 (Online)
ISBN 978-1-76016-282-5 (PDF)

LSAC_NAPLAN-Medicare_paper28

Contents

List of figures

List of tables

Glossary

Term	Description
LSAC	*Growing Up in Australia*: The Longitudinal Study of Australian Children
NAPLAN	The National Assessment Program – Literacy and Numeracy
Medicare	Australian Commonwealth Government-funded program that provides access to free or subsidised treatment under the Medicare Benefits Schedule
Observation window	Period of time in which all relevant measurements are taken
Duration effect	A form of measurement error. Study participants have observation windows of different duration. Those observed for longer periods of time have more opportunity to experience the event of interest than those observed for shorter periods.
Population coverage	The proportion of the population of interest that has information available in the sample

Overview

This paper highlights the potential of *Growing Up in Australia*: The Longitudinal Study of Australian Children (LSAC) to answer research questions using a combination of survey and linked administrative data. It focuses on two databases linked to LSAC: the Medicare Australia database and the National Assessment Program – Literacy and Numeracy (NAPLAN). Multisource data of this nature are complex and complicated to manipulate and analyse. Firstly, analysts need to resolve issues arising from different temporal patterns of data collection. LSAC surveys are administered every two years, NAPLAN occurs biyearly but not necessarily in the same year as LSAC surveys, and Medicare data are effectively collected in continuous time. These disparate periodicities make it difficult to align data points contemporaneously. Secondly, coverage of the population of interest varies across the three sources. Records could be missing from the longitudinal survey data due to Wave non-response, whereas missing NAPLAN and Medicare records arise not only from Wave non-response but also lack of consent to linkage, inability to match records during the linkage process, or students not sitting NAPLAN assessments, for example.

This paper demonstrates methods analysts can use to familiarise themselves with a multisource dataset comprised of LSAC survey, Medicare and NAPLAN data. We provide a straightforward example that researchers can use as a foundation or model that informs their own analysis. Specifically, the case study investigates whether Medicare service use mediates achievement in Year 9 NAPLAN reading scores among adolescents with a health condition. We show readers how to define an observation window – that is, the period of time in which all relevant measurements were taken – paying close attention to the upper and lower limits and underlying temporal patterns of data collection. We detail how the sample was selected and examine population coverage and bias in the multisource dataset.

Summary of key points

- Analysts working with a multisource dataset comprised of LSAC survey, Medicare and NAPLAN data need to understand the different temporal patterns of data collection in each. They will need to decide an appropriate common unit and metric for indicating the time of collection; possibilities include age measured in years or months, Wave of survey administration or calendar year, for example.
- Researchers can gain insight into the timing of observations, and consequences of decisions made on how to measure time, by graphing a subsample of cases and their data points on an axis that is labelled and marked in the chosen common unit.
- Analysts should clearly define an observation window, which is the period of time within which all measurements relevant to the analysis were taken.
- Analysis may need to account for either or both of duration or period effects. Duration effects occur where some individuals have longer observation windows than others, and hence a higher chance of experiencing the event of interest. Period effects can occur where measurements were taken at different periods in time. For example, Year 9 NAPLAN for LSAC K cohort occurred between 2012 and 2015.
- Researchers are advised to undertake appropriate analysis to determine the impact of using multisource linked administrative data on population coverage. That is, to gain an understanding of how many individuals are removed from the analytic sample due to missing data in each composite data source, the characteristics of these individuals and impact on sample representativeness.

Introduction

Growing Up in Australia: The Longitudinal Study of Australian Children (LSAC) is a major study following the development of 10,000 young people and their families. The study began in 2003 with a representative sample of children from urban and rural areas of all states and territories in Australia. It has a multi-disciplinary base and examines a broad range of research questions about development and wellbeing over the life course in relation to topics such as parenting, family, peers, education, child care and health. The study informs social policy and is used to identify opportunities for early intervention and prevention strategies. It follows two cohorts of children: the 'B' (baby) cohort were born between March 2003 and February 2004 and the 'K' (kindergarten) cohort were born between March 1999 and February 2000. This paper uses data from the K cohort, which had 4,983 responding families at Wave 1.

LSAC links to administrative databases (with participant consent) thereby adding valuable information to supplement, or complement, data collected during fieldwork. Administrative data can facilitate the study of social issues and policy change by providing information that would otherwise be unavailable, increasing opportunities for analysing outcomes and risk factors across different domains (Connelly, Playford, Gayle, & Dibben, 2016; Jutte, Roos, & Brownell, 2011). This paper focuses on two databases linked to LSAC: the Medicare Australia database and the National Assessment Program – Literacy and Numeracy (NAPLAN).[1] Medicare data contain details about participants' medical history and are considered accurate with regards to measuring access to health care services and benefits provided by the government (Parkinson, van Gool, & Kenny, 2011). NAPLAN occurs annually in Grades 3, 5, 7 and 9 and assesses academic performance across the following domains: reading, writing, conventions of language and numeracy (Daraganova, Edwards, & Sipthorp, 2013). While Medicare and NAPLAN enhance survey data, survey data, in turn, contributes valuable contextual information and insight to the administrative source (Connelly et al., 2016; Hand et al., 2018). For example, survey data can provide details on individuals' family and social context and wellbeing prior to, during or after engagement with health care services and/or educational testing.

To date, there is limited published research that has utilised all three sources of the LSAC survey, Medicare and NAPLAN simultaneously. This may be, at least in part, due to the complexities involved. Research data literacy is defined as the set of skills and knowledge needed to transform data into information and actionable knowledge (Koltay, 2016), and it can be difficult to achieve with multisource datasets. Administrative data are not designed or collected for the purpose of research and challenges can arise, for example, where it is collected with different periodicities or where coverage of the population of interest is not complete, impacting the validity of inferential statements (Hand et al., 2018). The timing and frequency of data collection for each of the LSAC survey, Medicare and NAPLAN are different and it is difficult for analysts to align observations contemporaneously: surveys are distributed every two years, while NAPLAN assessments happen every two years but not in the same calendar year for all children and not necessarily in the same year as surveys are administered. Medicare data are collected in continuous time, but patterns of use are very diverse. Some children access services multiple times per year while others have very occasional use, if at all.

It is often thought that administrative data cover whole populations of interest (Connelly et al., 2016). However, that isn't necessarily true. Both administrative and survey data are rarely complete and it is recommended that analysts undertake work to understand missingness and non-representativeness, particularly where the aim is to make inference to another population (Hand et al., 2018). Missing records in LSAC stem from non-participation, whereas missing cases in Medicare or NAPLAN could arise from non-participation, lack of consent for linkage or inability to match records to an individual during the linkage process, for example. Thus, population coverage can vary across the three datasets and as Hand and colleagues (2018) advise, 'quality issues of individual databases may propagate and amplify', potentially impacting on the accuracy and validity of conclusions (Agafiței, Gras, Kloek, & Reis, 2015).

[1] Other databases linked to LSAC include Centrelink Welfare; Australian Childhood Immunisation Register; Australian Early Development Census; MySchool; Pharmaceutical Benefits Scheme. Further details are available at growingupinaustralia.gov.au/data-and-documentation/lsac-data-linkages.

Aim of this paper

The aim of this paper is to improve and contribute to research data literacy by showing how to approach and manipulate a multisource dataset comprised of LSAC survey, Medicare and NAPLAN data. Readers are guided through methodological issues relating to periodicity, timing and alignment of data points, as well as population coverage and assessing sample bias.

We illustrate via a case study approach; this is beneficial when there is a need to appreciate an issue or phenomena in a real-life context (Crowe et al., 2011). The case study investigates whether the use of Medicare services mediates the impact that a self-reported health condition might have on year 9 NAPLAN reading levels. There is limited research examining educational outcomes for Australian adolescents with chronic health conditions (Jackson, 2013) but analysis using Growing Up in Ireland (part of the 'Growing Up' series along with LSAC) showed that adolescent chronic illness can have significant negative effects on standardised reading test scores (Layte & McCrory, 2013). For the purposes of this paper, a health condition was defined as any medical condition or disability that had lasted or was likely to last for 6 months or more, as reported by LSAC study children.

Note that this paper does not present a fully developed theoretical framework or research model; rather, we describe a relatively simple example that is sufficient for illustrating the methodological issues at hand and which analysts can use as a template to develop their own more complex analysis and models. The processes and strategies detailed for the concurrent analysis of survey, Medicare and NAPLAN data are applicable to literacy and numeracy outcomes as well as reading and could be adapted for research into specific health conditions rather than the general measure used here.

The next section introduces the case study and its research questions. We then discuss key methodological considerations around data sources and measures, including operationalising reading attainment, health status and Medicare use, defining an observation window (the period of time in which all relevant measurements were taken) and understanding temporal patterns of data collection. Issues relating to sample selection and population coverage are considered following that. Results for the case study and a summary conclude the paper.

Case study: Research questions

The research question of interest is whether the use of Medicare services mediates the impact of a health condition on reading levels at year 9. Using data from LSAC K cohort (*N* = 3,206), this was addressed using a mediation model (MacKinnon, Fairchild, & Fritz, 2007), constructed in three stages, that determined:

1. Do reading levels differ between adolescents with a health condition and those without?
2. Does the use of Medicare differ between adolescents with a health condition and those without?
3. Does the magnitude of any effect of health on reading levels lessen after accounting for use of Medicare services?

Methodological consideration 1: Data sources and measures

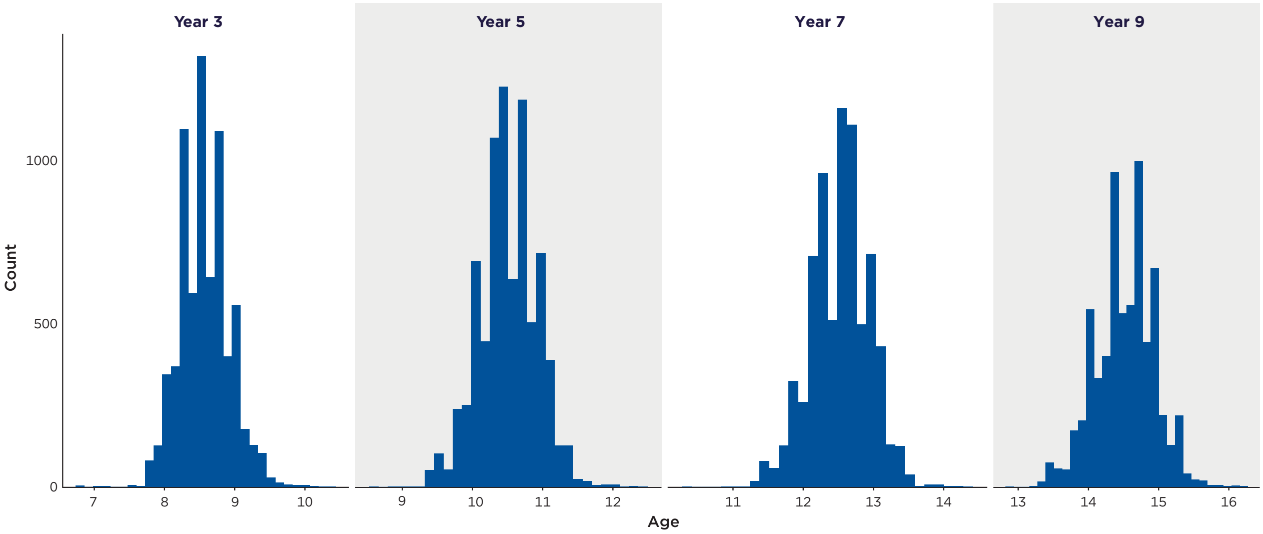
The first set of methodological considerations relate to data sources and operationalising the measures of interest, namely reading attainment, health status and use of Medicare services. We began by identifying the outcome variable – NAPLAN reading scores, in this case – and familiarised ourselves with when those assessments were taken. Then, we defined an observation window that used the timing of reading assessments as the foundation or cornerstone. The structure of the observation window and time points that defined it, in turn, informed the operationalisation of health status and use of Medicare. This section details this process as well as the related issue of duration effects.

NAPLAN reading scores

The NAPLAN dataset (**lsacnaplan**) provides information on children's year 9 reading level. It was a continuous scaled score given by the variable **y9read**. Scores were provided to one decimal point unless state or territory authorities provided scores rounded to the whole number. Scores typically ranged from 0 to 1,000, where a higher score meant a higher level of achievement.

NAPLAN assessments for K cohort children took place in school years 3, 5, 7 and 9 but children were not all the same age when assessed. Variables **y3age**, **y5age**, **y7age** and **y9age** in the NAPLAN dataset give the age of the child at time of assessment (the most recent assessment for those who sat twice due to repeating a year). Figure 1 shows the distribution of ages per school year. At the time of year 3 NAPLAN children's ages ranged from 6.8 to 10.4 years; at the year 5 assessment they ranged from 8.6 to 12.4 years; year 7 from 10.3 to 14.3, and in year 9, which is relevant to our case study, from 12.8 to 16.2 years.

Figure 1: Distribution of study child age at time of NAPLAN assessment by school year, LSAC K cohort



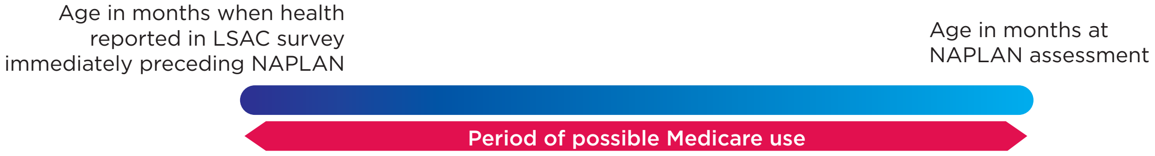
Defining the observation window

We needed to define an appropriate time interval, or observation window, relevant to the research question to be able to identify and align students' reported health status and Medicare usage contemporaneously with their NAPLAN reading scores. For the mediation model, health status had to be measured prior to a student sitting NAPLAN and Medicare use needed to be evaluated between the time of reported health status and NAPLAN.

A key decision was to choose a common unit of time for indicating when data collection took place in each source. Calendar year, month and/or day was one option, as was age in years, months and/or days. Age in months was the most sensible choice because it allowed for a more meaningful narrative and reading scores were more likely to be responsive to a student's age than to the calendar year.[2] For further information on issues to consider when choosing a metric for time see Singer and Willett (2003).

The age at the time of an individual's year 9 NAPLAN assessment identified the closing point of their observation window (Figure 2). As described above, this was given by variable **y9age** and values ranged from 12.8 to 16.2 years.

Figure 2: Observation window


Age in months when health reported in LSAC survey immediately preceding NAPLAN
Age in months at NAPLAN assessment
Period of possible Medicare use

[2] Analysts choosing calendar time for their project might find the following useful when deriving dates: In LSAC survey data ***datint** gives interview dates in day, month and year. Note, all days are given as the first of the month due to confidentialisation. NAPLAN contains calendar year of assessment in: **y3test**, **y5test**, **y7test**, **y9test**, **ry3test**, **ry5test**, **ry7test**, **ry9test**. Day, month and year of Medicare service and processing are given by **dateserv** and **datepay** respectively.

The observation window opened at the age at which each sample member reported their health status in the LSAC Wave that immediately preceded their year 9 NAPLAN assessment. LSAC surveys were administered every two years with Wave 1 for the K cohort taking place in 2004/5. Age at the time of each interview, to the nearest month, was derived using date of birth from Wave 1 (**zf04m1**) and date of interview in each Wave (***datint**). Figure 3 shows the age distribution at time of interview for respondents by wave. Wave 1 respondents were aged between 4 and 5.7 years, Wave 2 between 5.9 and 7.9 years and so on, with ages at Wave 8 between 18.2 and 20.1 years. Note that 329 individuals from Wave 8 were not included in the Figure because their date of interview was not available.

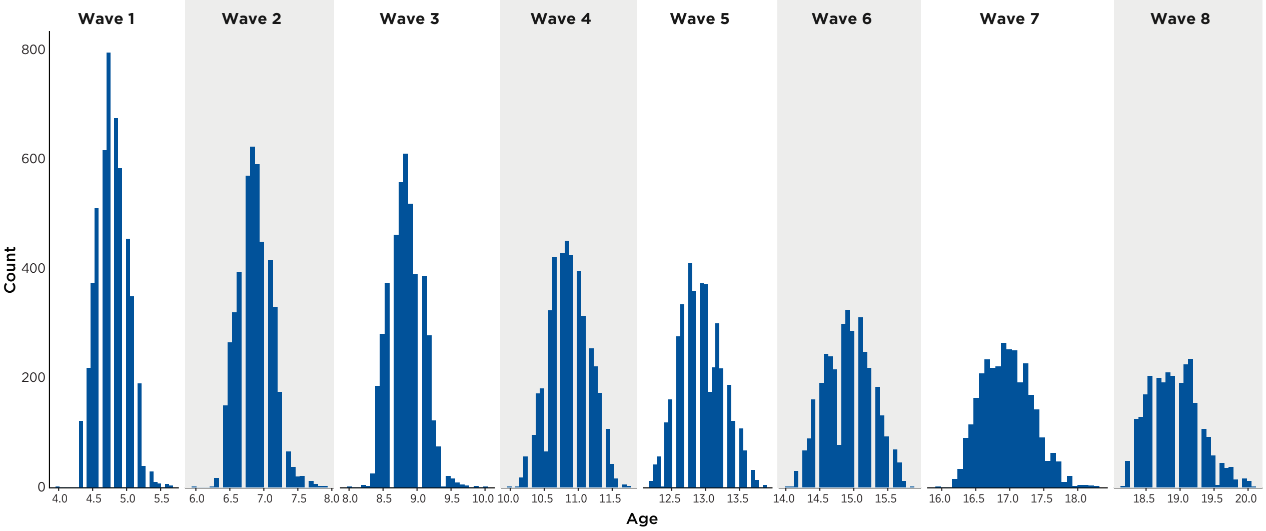
The observation window for analysis of year 9 reading scores ended at ages 12.8–16.2; therefore, the health measurement for any individual in our study could be sourced from survey Waves 5, 6 or 7.

Operationalising health status

LSAC survey Waves 5, 6 and 7 were used to identify children who reported having a health condition and those who reported having no such condition (using item ***f13m1**). Age at the time of each survey interview and age at time of the NAPLAN assessment (both rounded to the nearest month) were compared to identify the health status that immediately preceded the assessment. A binary indicator was derived using that status, that took value 1 = had medical condition or 0 = no health condition. Status was taken from Wave 5 for 88% of the sample and Wave 6 for the remainder, with no observations for this particular sample taken from Wave 7.

In both Waves 5 and 6 the survey question asked, ‘Does the family member (that is, the study child) have any medical conditions or disabilities that have lasted, or are likely to last, for six months or more?’ Conditions or disabilities captured in this measure could have included hearing, speech and sight problems (not corrected by glasses), learning difficulties, mental illness, chronic pain, nervous conditions, limited use of arms, fingers, legs or feet, fits or loss of consciousness or any other long-term conditions.

Figure 3: Distribution of study child age at time of interview by wave, LSAC K cohort



Deriving the Medicare use variable

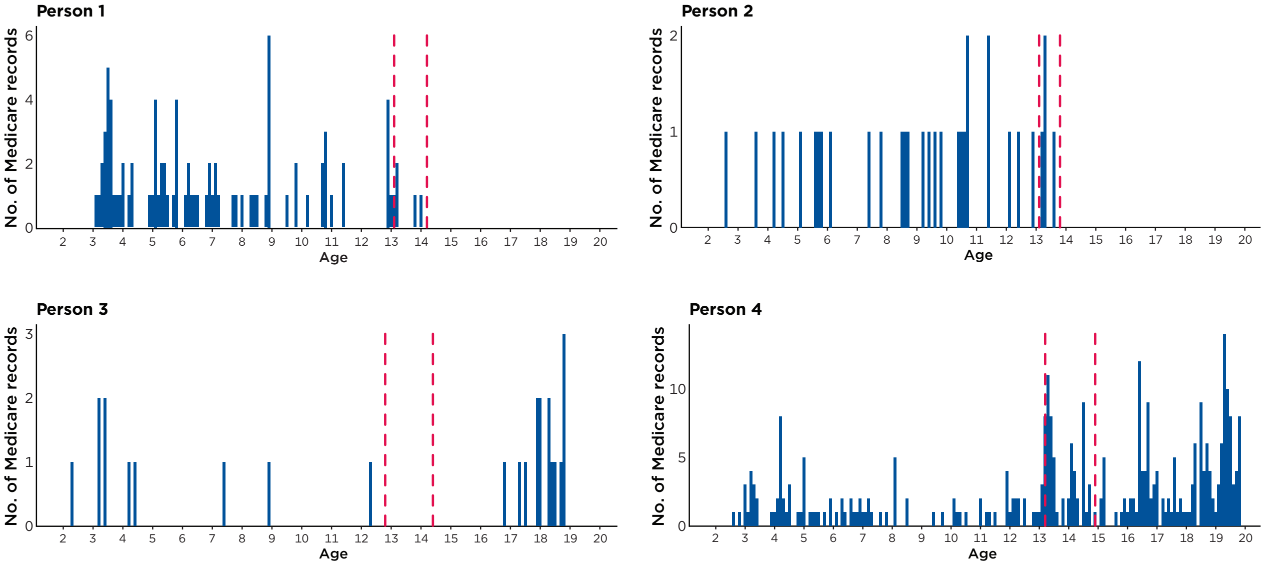
Our research questions concerned Medicare use within individuals’ observation windows as defined above; that is, between sitting NAPLAN and the time that their health status was known prior to that, where time was measured using age. Records in the Medicare dataset (**mbssc**) contain date of service (**dateserv**) and date of processing (**datepay**). The child’s age at the time of service use was derived by merging date of birth information from Wave 1 survey data (**zf04m1**) with the Medicare dataset using the unique **hicid** identifier. Age in months was calculated as the difference between this and date of service or, if missing, date of processing. One hundred and sixty-nine records had no date details at all. These records were removed from the analysis, as without date information it was not possible to align them with NAPLAN and LSAC survey data and include them in research that required chronological ordering of events and circumstances. This did not result in any individual people being removed from the dataset.

We counted the number of Medicare records within each individual's observation window and constructed a binary indicator, with value = 1 if an individual had at least one Medicare record in that time and 0 otherwise. Note that the availability of a record did not necessarily signify a complete record. With the exception of records missing date information, those lacking other details (such as medical service used or benefit amount) were included in this analysis. Thus, counts of records and references to them included those with partial or incomplete information.

Figure 4 shows the number of Medicare records documented by age (age rounded to one decimal place) for four members from the sample chosen for illustrative purposes. The two dashed vertical lines on each graph indicate the observation window relevant to that individual. Person 1 had five records in their observation window; Person 2 had four; Person 3 had none and Person 4 had 63. The Figure illustrates the extent to which Medicare information has been compressed for this case study, in that the binary indicator masks the exact number of services used and precise details of when the use happened. We have deliberately kept this indicator simple for the purposes of this paper, but readers are encouraged to consider other forms that capture more of the information available about Medicare service use as appropriate for their research (e.g., a categorical variable that indicated low, medium or high use, or differentiated by service type, or a measure of financial cost).

There may be a small percentage of individuals classified as non-users of Medicare when they had used a service, and the count of records could be underestimated for some. This could occur where an individual had used Medicare services but no records were present in the **mbssc** dataset, due to not giving consent to linkage; non-participation in an LSAC wave; or inability to match a record to an LSAC study child during the linkage process. Work is currently underway to determine if information on individual consent could be provided in future data releases.

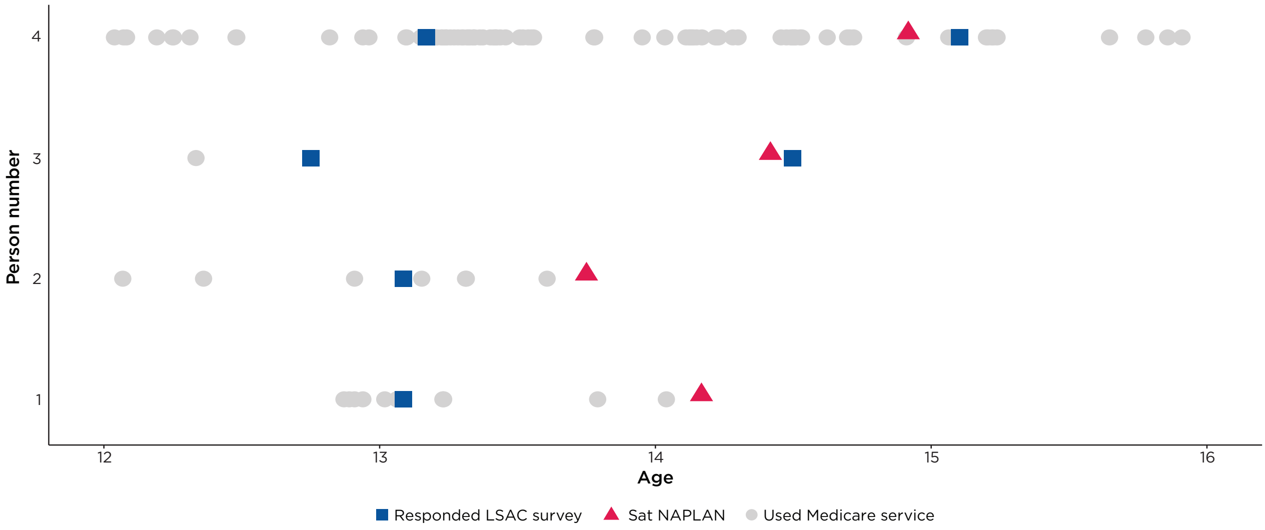
Figure 4: Distribution of Medicare records for four illustrative cases



Understanding duration effects

Plotting datapoints for a small number of cases can aid understanding of temporal patterns of data collection and the consequences of decisions made when operationalising variables. Figure 5 shows ages at which survey interviews were conducted, at least one Medicare service was used, and year 9 NAPLAN tests were sat for the four illustrative cases used above. The horizontal axis is age calculated in months, truncated from 12 to 16 years. Each blue square signifies a response to the survey item ***f13m1**, which indicated whether the study child had any medical conditions or disabilities that lasted or were likely to last for six months or more. Each red triangle shows the age at which the respondent sat NAPLAN in year 9 and hence marks the end of the observation window. Grey dots indicate when at least one Medicare service was used (i.e., record was available).

Figure 5: Timing of LSAC survey, Medicare and Year 9 NAPLAN observations for four illustrative cases



To illustrate, Person 1 reported their health status just after turning 13 years of age (at 13.1 years). They sat year 9 NAPLAN a few months after their fourteenth birthday (at 14.2 years). Their observation window for this analysis therefore opened at 13.1 and closed at 14.2 years. Medicare service use occurred in the six months prior to reporting their health status and was infrequently used within the observation window. Person 2 reported on their health at the same age as Person 1 (13.1 years) but were younger when they sat NAPLAN (13.8 years). Thus, their observation window was shorter (by around 4 months) than that for Person 1.

Person 3 sat year 9 NAPLAN when they were 14.4 years old. Their health status was reported when they were 12.8 and again at 14.5 years. For the purposes of our analysis, their observation window was taken as 12.8–14.4 years; it opened at the younger age of 12.8 due to the requirement that the LSAC survey occurred prior to sitting the NAPLAN assessment. Similarly, the observation window for Person 4 was from 13.2 years, when their health status was reported prior to NAPLAN, and 14.9 years when NAPLAN was sat.

The above case studies highlight possible duration effects that need to be acknowledged and accounted for in the analysis. The duration of the observation windows for Persons 1–4 are 1.1 years; 0.7 years; 1.6 years and 1.7 years, respectively. Thus, Persons 1 and 2 had a shorter period of observed time in which to utilise Medicare services, compared to Person 3 and 4. Period effects may also occur where measurements were taken at different calendar years or periods of time. For example, year 9 NAPLAN for the LSAC K cohort occurred between 2012 and 2015.

Methodological consideration 2: Sample selection and population coverage

This section details how the analytic sample was identified from the LSAC survey, Medicare and NAPLAN datasets; characteristics of individuals in that multisource sample, and how it compared to the original Wave 1 sample, which was considered broadly representative.

Sample selection

The population of interest in this case study was year 9 students with a health condition. The analytic sample was identified as follows:

1. Each case in the NAPLAN dataset represented an LSAC study child recruited at Wave 1 (*N* = 4,983 for K cohort). The variable **consent** identified those with consent to linkage of NAPLAN data (*N* = 4,227).
2. Of those who consented, 582 were missing year 9 reading scores and were removed from the sample, giving *N* = 3,645. Eight of these 582 occurred where LSAC study members couldn't be identified in the NAPLAN database by state and territorial authorities responsible for matching (indicated by the variable **y9status**). The most likely explanation for missing reading scores for the other 574 individuals was sample attrition.
3. A further 243 individuals were removed as they were either absent, exempt or withdrawn from the Year 9 reading assessment (shown by **y9read** and **y9status**), giving *N* = 3,402.
4. 168 individuals were removed because their age at the time of measuring health immediately preceding NAPLAN couldn't be determined (using date of interview ***datint** and date of birth **zfo4m1**), giving *N* = 3,234.
5. Finally, 28 individuals were removed because their health status was not reported. The final sample size was *N* = 3,206. Note the sample included six individuals who repeated year 9 (indicated by variable **ry9**); if they sat NAPLAN for a second time, their most recent score was used.

The above process focused on the removal of cases due to missing data in survey and NAPLAN data, not Medicare. For this particular study, we assumed the remaining 3,206 cases had consented to Medicare linkage and their records were able to be matched. Consequently, no further cases were removed. Currently, the NAPLAN dataset contains a consent variable (indicating whether participants had consented to linkage). However, an equivalent variable is not available in the Medicare dataset (work is underway to determine if this is possible and could be provided in future data releases).

Population coverage

We studied the characteristics of individuals in both the original (*N* = 4,983) and analytic sample (*N* = 3,206) to examine representativeness (Table 1). The analytic sample had 1,777 fewer individuals, who were lost or removed due to a combination of attrition, lack of consent to NAPLAN linkage, unknown year 9 reading scores, and missing date or health information. Results showed that while the two samples were similar in composition in some respects, there were differences in others.

The Wave 1 sample was considered broadly representative of the Australian population born between March 1999 and February 2000 (Gray & Sanson, 2005; Mohal et al., 2021). The final analytic sample had a similar proportion of females as the Wave 1 sample (49.2% vs 49.1%) and study children born in Australia (96.0% vs 95.8%). It contained a slightly higher proportion of children with mothers born in Australia (64.4% vs 61.1%) and mothers who had completed year 12 or equivalent (64.8% vs 58.1%) but a lower proportion of children of Aboriginal or Torres Strait Islander origin (2.4% vs 3.8%).

Table 1: Sample composition

	Sample at Wave 1		Final analytic sample	
	n	%	*n*	%
Female	2,447	49.1	1,576	49.2
Study child born in Australia	4,772	95.8	3,079	96.0
Mother born in Australia	3,046	61.1	2,066	64.4
Study child of Aboriginal or Torres Strait Islander origin	187	3.8	76	2.4
Mother completed year 12 or equivalent	2,895	58.1	2,077	64.8
Total sample size	4,983		3,206	

Case study: results

Over 3% of individuals (3.5%, N = 111) reported a health condition. Seventy-nine percent (N = 2,538) had used Medicare between reporting on their health status and sitting their NAPLAN assessment. Reading scores ranged from 195.6 to 907.5, with a mean of 599.2 (Table 2).

There was some variation in the sample in the age of respondents when health status was reported, and in the length of observation windows. Age at time of reporting ranged from 12.16 years to 15.34 years, with a mean of 13.15 years. Observation windows varied from one month to around three years, meaning that duration effects could at least partly explain variance in reading scores as some individuals had more time than others to access health care and for other events to occur that could impact reading scores. Reading scores might also be subject to period effects. Around one-quarter of sample members (24.4%) sat their NAPLAN assessment in 2013, 71% in 2014 and 4.5% in 2015.

Table 2: Summary statistics of analytic sample

Statistic	Value
Mean age when health measurement taken	13.15 years
Range of ages when health measurement taken	12.16–15.34 years
Mean length of observation window (number of years between health measurement and NAPLAN)	1.31
Range of number of years between health measurement and NAPLAN	0.0002–3.00
Mean reading score	599.2
Range of reading scores	195.6–907.5
Range of calendar years when NAPLAN was completed	2013–2015

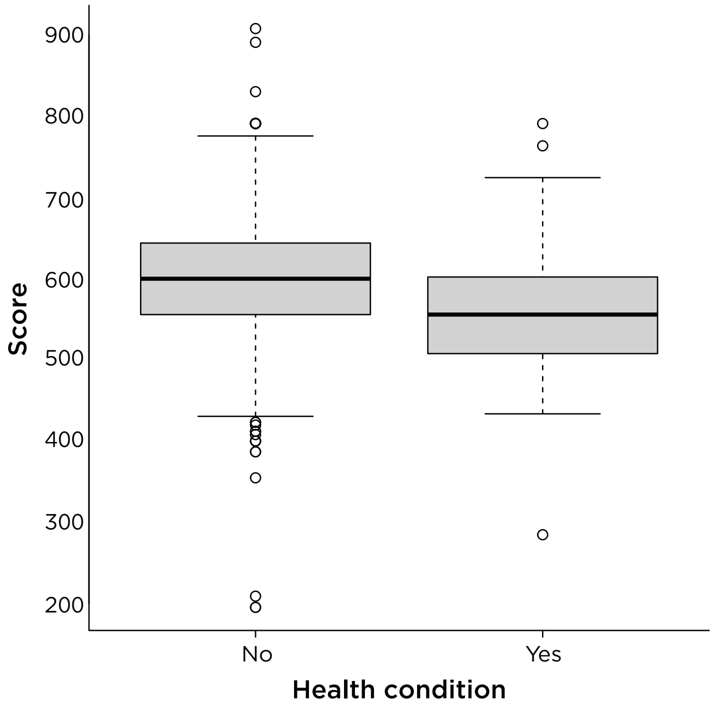
Do reading scores differ between adolescents with health conditions and those without?

The distributions of reading scores by health status are summarised in Table 3 and Figure 6. The mean reading score of students with a health condition was notably lower than it was for those with no health condition (559.5 compared to 600.6). This difference was statistically significant at the 5% level (Welch's t-test; t_{117} = 5.683, p < 0.001). The students with no health condition included several with reading scores that were very high or very low (i.e. outlying values, defined as those more or less than 1.5 times the interquartile range); there were fewer such values observed among students with health conditions, leading to a lower range of scores (505.8 compared to 711.9). The higher standard deviation in students with health conditions compared to none (75.1 vs 68.4) reflected the relatively small sample size (N = 111 vs N = 3,095).

Table 3: Descriptive statistics for reading score by health condition

	No health condition	Has health condition
Mean	600.6	559.5
Minimum	195.6	284.9
Maximum	907.5	790.7
Range	711.9	505.8
Standard deviation	68.4	75.1
N	3,095	111

Figure 6: Distribution of year 9 NAPLAN reading scores by health condition



Does use of Medicare differ between adolescents with a health condition and those without?

The proportion of adolescents who used Medicare services was higher among those who reported having a health condition than among those with no such condition (Table 4). Ninety percent of individuals with a health condition had used a Medicare service at least once, compared to 78.8% of those with no reported condition. Chi-square tests confirmed an association between health status and Medicare use in both samples $(\chi^2_1 = 8.321, p = 0.004)$.

Table 4: Number and proportion of adolescents who used Medicare services on at least one occasion in observation window

	Used Medicare		Did not use Medicare		Total	
	n	%	*n*	%	*n*	%
No health condition	2,438	78.8	657	21.2	3,095	100.0
Has health condition	100	90.1	11	9.9	111	100.0

Does use of Medicare mediate impact of health condition on NAPLAN reading scores?

Three models were used to examine the possible mediating effect of Medicare access on year 9 reading scores among children with health conditions. Details of model specifications and results follow. Unstandardised regression coefficients are reported throughout (symbolised by $\hat{\beta}$, to represent the estimated population parameter β).

The models below have been adjusted for duration effects as described, but do not account for possible period effects. Members of the sample sat NAPLAN in different calendar years and health measurements were also taken at different times. It wasn't possible to adjust models for these differences due to the relationship between calendar year and other measures; the year NAPLAN was sat, for example, was correlated with the duration indictor. Issues such as this can be raised and discussed as limitations to analysis and research findings.

Model 1: Impact of health on reading scores

Model 1 was a linear regression model with outcome of reading score in continuous form. The predictor variable was a binary indicator of whether the individual had a health condition or not (0 = no health condition, 1 = had health condition). The model also included a continuous variable for the number of years between reporting of health and sitting the NAPLAN assessment; that is, an adjustment for possible duration effect.

Results are in Table 5. They indicate a statistically significant impact of having a health condition on reading scores at year 9 when assessed at the 5% level $(\hat{\beta} = -41.73, p < 0.001)$. Children with a health condition would, on average, be expected to have reading scores around 42 points lower than their peers with no such condition.

Table 5: Results of linear regression model for predicting NAPLAN reading score

	Coefficient	*SE*	*p*
Intercept	592.16	2.93	<0.001
Has health condition (ref: no)	-41.73	6.62	<0.001
Number years between health report and NAPLAN assessment	6.45	2.03	0.002
Calendar year NAPLAN assessment	1.19	2.64	

Notes: *SE* = standard error.

Model 2: Impact of health condition on Medicare use

Model 2 was a logistic regression with the outcome measure of whether Medicare services were used or not (0 = no health condition, 1 = had health condition). The model controlled for the number of years between health reporting and NAPLAN assessment (i.e. the duration effect). Results are in Table 6. They showed that among students who sat NAPLAN in year 9, those with a health condition were 2.41 times more likely to use Medicare than those with no condition and this was significant at the 5% level $(exp(0.88) = 2.4109, p = 0.012)$.

Table 6: Results of logistic regression model for predicting use of Medicare services

	Coefficient (log odds)	*SE*	*p*
Intercept	-0.83	0.10	<0.001
Has health condition (ref: no)	0.88	0.35	0.012
Number years between health report and NAPLAN assessment	1.88	0.08	<0.001

Notes: *SE* = standard error.

Model 3: Impact of health on reading score accounting for Medicare use

The third specification was a linear regression with outcome of reading score. The primary aim was to assess whether the relationship between having a health condition and reading score established in Model 1 changed after accounting for Medicare use. That is, whether use of Medicare mediated the effect of health on reading attainment. The outcome in the linear regression was continuous reading score and predictor variables were presence of a health condition, use of Medicare, and number of years between reporting health and sitting NAPLAN.

As shown above, prior to accounting for Medicare use, students at year 9 with a health condition had reading scores that were 41.73 points lower than those for students without a health condition. Results of the mediation model in Table 7 showed a slightly lower estimated effect of 42.05 points of having a health condition after accounting for Medicare use $(\hat{\beta} = -42.05, p < 0.001)$. Having accessed any Medicare service on at least one occasion had a limited effect on reading scores among children with health conditions; the effect persisted after controlling for Medicare use.

The above findings shouldn't be regarded as definitive. As stated earlier, the aim of this paper wasn't to present a comprehensive analysis and model. Rather, we have described an elementary model containing basic indicators and measures; the indicator of Medicare use, in particular, was rudimentary, and this may explain the results found.

Table 7: Results of linear regression mediation model with outcome of NAPLAN reading score

	Coefficient	*SE*	*p*
Intercept	590.68	3.20	<0.001
Has health condition (ref: no)	-42.05	6.63	<0.001
Used Medicare (ref: no)	3.86	3.34	0.249
Number years between health report and NAPLAN assessment	5.26	2.28	0.021

Notes: *SE* = standard error.

Summary

The case study presented in this technical report examined the possible mediating effect of Medicare use on year 9 NAPLAN reading scores among adolescents with self-reported health conditions. The paper discussed two particular aspects of working with combined LSAC survey and linked NAPLAN and Medicare datasets. Firstly, the measurement of time and different temporal patterns of data collection and, secondly, sample selection and population coverage.

Compiling one single dataset from LSAC survey and multiple linked sources requires integrating records collected at disparate points in time. We showed how to define and construct an observation window, defined as the period of time in which all relevant measurements were taken. For our case study, it began at the age in months when respondents reported their health condition in the LSAC Wave that immediately preceded their year 9 NAPLAN reading assessment. It closed at the age of that assessment, and an indicator of Medicare use was derived for the period of time in-between those points. The length of time, or duration, of observation windows varied across the sample and this was accounted for in the modelling process. This was important as some sample members had a longer period of time in which they could have accessed Medicare compared with others.

Administrative data should, in theory, cover whole populations of interest (Connelly et al., 2016). That isn't always the case, however. Representation in the LSAC-linked Medicare and NAPLAN datasets is contingent on several factors including consent to linkage, participation in relevant survey waves and ability to match records to an individual during the linkage process. Any under or over representation of a population subgroup in each individual data source could be exacerbated when combined and integrated with survey data. In our case study, males and females, and study children born in Australia were appropriately represented in the final composite dataset. There were fewer than expected with mothers who had not completed year 12 and those of Aboriginal or Torres Strait Islander origin, indicating less presence and coverage of these population groups in our analysis.

The case study presented in this paper showed a limited mediating effect of Medicare use on year 9 NAPLAN reading scores among adolescents with a health condition. Evidence suggests that students with health conditions had significantly lower scores than those with no condition, and they also had more occasions where they accessed at least one Medicare service. In the basic model detailed here, Medicare use did not alter the effect of having a health condition on reading scores in a meaningful way. However, as mentioned throughout, the analysis and model constructed in the case study was not comprehensive and results should be viewed as tentative. We hope that researchers are motivated to develop their own work using LSAC survey, NAPLAN and Medicare data, using the insight into timing and coverage issues provided by this paper.

References

Agafiței, M., Gras, F., Kloek, W., & Reis, F. (2015). Measuring output quality for multisource statistics in official statistics: Some directions. *Statistical Journal of the IAOS, 31(2)*, 203–211.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12.

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., & Sheikh, A. (2011). The case study approach. *BMC Medical Research Methodology, 11*(1), 1–9.

Daraganova, G., Edwards, B., & Sipthorp, M. (2013). *Using National Assessment Program Literacy and Numeracy (NAPLAN) data in the Longitudinal Study of Australian Children (LSAC)*. Melbourne: Australian Institute of Family Studies. Retrieved from www.growingupinaustralia.gov.au/sites/default/files/tp8.pdf

Gray, M., & Sanson, A. (2005). *Growing Up in Australia*: The longitudinal study of Australian children. *Family Matters, 72*, 4–9.

Hand, D. J., Babb, P., Zhang, L.-C., Allin, P., Wallgren, A., Wallgren, B. et al. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 181*(3), 555–605.

Jackson, M. (2013). The special educational needs of adolescents living with chronic illness: A literature review. *International Journal of Inclusive Education, 17*(6), 543–554. doi:10.1080/13603116.2012.676085

Jutte, D. P., Roos, L. L., & Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual Review of Public Health, 32*, 91–108.

Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA Journal, 42*(4), 303–312.

Layte, R., & McCrory, C. (2013). Paediatric chronic illness and educational failure: The role of emotional and behavioural problems. *Social Psychiatry and Psychiatric Epidemiology, 48*(8), 1307–1316. doi:10.1007/s00127-012-0609-3

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*(1), 593–614. doi:10.1146/annurev.psych.58.110405.085542

Mohal, J., Lansangan, C., Gasser, C., Taylor, T., Renda, J., Jessup, K. et al. (2021). Growing Up in Australia: *The Longitudinal Study of Australian Children–Data User Guide, Release 9C1, June 2021*. Melbourne: Australian Institute of Family Studies.

Parkinson, B., van Gool, K., & Kenny, P. (2011). *Medicare Australia data for research: An introduction*. Retrieved from www.crest.uts.edu.au/pdfs/Factsheet-Medicare_Australia-FINAL.pdf

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.