International Journal of Population Data Science

Journal Website: www.ijpds.org

Data resource profile: Education and Health Data in New Zealand's Integrated Data Infrastructure

Thomas Schober[1,*], Nicholas Bowden[2], Stephanie D'Souza[3], Sheree Gibb[4], Barry Milne[3], and Lisa Meehan[1]

Submission History

Submitted:	30/09/2025
Accepted:	05/02/2026
Published:	15/04/2026

[1] New Zealand Policy Research Institute, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand
[2] Department of Paediatrics and Child Health, University of Otago, PO Box 56, Dunedin 9054, New Zealand
[3] Centre of Methods and Policy Application in the Social Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[4] Department of Public Health, University of Otago, PO Box 7343, Newtown, Wellington 6242, New Zealand

Abstract

This paper provides an overview of education and health data in New Zealand's Integrated Data Infrastructure (IDI). The IDI combines population-wide administrative, census, and survey data from government agencies and other organisations, enabling detailed research into the interplay between education and health at the individual level. We describe the structure of the IDI, key data sources in education and health, and provide an overview of access procedures and practical considerations for working with the data.

Keywords

administrative data; education; health; New Zealand; Integrated Data Infrastructure (IDI)

Key features

- The Integrated Data Infrastructure (IDI) is a whole-population research database in New Zealand that links administrative, census, and survey data at the individual level.
- Administrative education records span early childhood through tertiary education, with information on enrolment, attendance, and attainment.
- Administrative health data include hospital discharges, pharmaceutical dispensing, mental health service use, immunisation records, maternity and mortality data.
- Additional sources include census and survey data that complement the administrative records.

*Corresponding Author:
Email Address: thomas-schober@aut.ac.nz (Thomas Schober)

https://doi.org/10.23889/ijpds.v11i2.3364

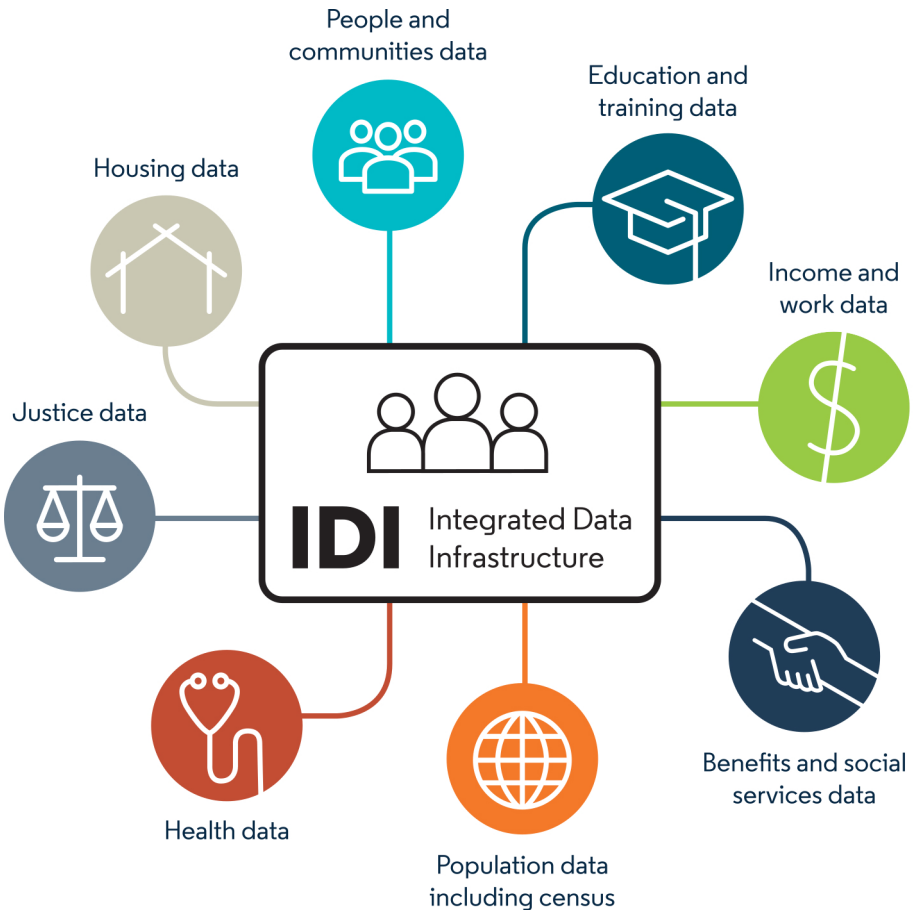
Introduction

The Integrated Data Infrastructure (IDI) is a cornerstone of empirical research and policy evaluation in Aotearoa New Zealand. Managed by Stats NZ, the country's national statistics office, the IDI is a large research database that holds microdata from a wide range of government agencies, organisations, and surveys [1]. All data are recorded and linked at the individual level, enabling researchers to examine social, economic, and health outcomes over time and across domains. Figure 1 provides an overview of the breadth of linked data available in the IDI, with each domain, such as education and health data, including a number of different datasets. Given New Zealand's resident population of more than 5 million people, including around 1.3 million children and young people under the age of 20 [2], administrative education and health data in the IDI support analyses based on near-complete population cohorts for their respective target populations. The IDI is widely used by academic researchers, government analysts, policy advisors, and independent research organisations. A regularly updated publication database lists over 1600 outputs (as of January 2026), including journal articles, government agency reports, reports from universities and research institutions, student theses, and non-empirical publications that focus on the IDI itself [3].

Education and health are two of the most comprehensively represented domains in the IDI. Administrative data from the New Zealand Ministry of Education (MOE) cover early childhood (ECE) through to tertiary education, capturing detailed records on enrolment, attendance, and attainment. These are complemented by a rich array of health data sourced from Health New Zealand - Te Whatu Ora (government health service delivery agency) and the Ministry of Health - Manatū Hauora (MOH).

The ability to link education and health data at the individual level enables research into how experiences or characteristics in one domain may influence outcomes in the other. Studies have used this potential to explore the association between health conditions, such as type 1 diabetes, autism, obsessive-compulsive disorder, anxiety, and depression, and educational outcomes such as achievement, attendance, exclusions and university enrolment [5–7]. Others have combined school assessment data with administrative records to examine how literacy and numeracy skills relate to later health and labour market outcomes [8], explored how education attainment is associated with outcomes across the lifecourse [9], utilised health and development screening assessment information to assess their predictive properties on future literacy interventions [10], and investigated how educational supports are associated with improved educational outcomes among students with significant health-related

Figure 1: Overview of IDI data



Source: Stats NZ [4].

need [11]. Such cross-sector analyses illustrate the value of integrated data for understanding lifecourse processes that cut across institutional boundaries.

This paper provides an overview of education and health data in the IDI, describing key administrative sources alongside linked census and survey data. It also outlines the structure of the IDI and explains how researchers can access and work with the data.

Background

The IDI was established in 2011 by Stats NZ, building on earlier experience linking employer and employee data. It emerged as part of a broader government effort to make better use of available administrative data, understand the impact of policies, and create evidence-based advice to support decision-making [12].

Stats NZ controls access to the IDI to ensure data security, privacy, and ethical use. Access is granted only when all conditions of the Five Safes framework are met, which manages risk across five domains: Safe Projects (appropriate use), Safe People (researchers are trusted and competent), Safe Data (data are treated and protected to minimise disclosure risk), Safe Settings (secure environments are used for access), and Safe Outputs (results are checked to prevent identification) [13]. Researchers must submit a detailed project proposal demonstrating public benefit and are vetted and required to complete mandatory training before access is approved. Access is restricted to secure Data Labs, physical rooms where only approved researchers can work with IDI data, and all datasets are de-identified, with identifying information removed. Finally, all outputs are checked before release to ensure confidentiality is maintained. At present, access is limited to secure Data Labs located in New Zealand. In addition to the Five Safes, research using the IDI must also respect the cultural context of New Zealand. Stats NZ has developed the *Ngā Tikanga Paihere* framework, which draws on values and principles of Māori, the Indigenous people of New Zealand, to guide safe and appropriate use of data and to ensure use of data serves the people and communities behind the data [12].

At the core of the IDI is a central 'spine' that aims to include all individuals who have ever resided in New Zealand. The spine is constructed using birth records, visa approvals, and tax data [14]. Other data collections are then linked to this spine using probabilistic record linkage. Datasets within the same sector may be linked using a common identification number (for example hospitalisations and pharmaceutical dispensing which are both within the Health collection and can be linked using National Health Index (NHI) number). However, New Zealand does not have a single universal personal identifier to allow for direct linking between different sectors, so all other linkage is probabilistic based on names, date of birth, and address. Before the data are made available to researchers, all identifying information is removed, so that individual records cannot be associated with named individuals. In addition, while some administrative datasets contain free-text fields, these are generally removed before being incorporated into the IDI to protect confidentiality. As a result, most variables available for analysis are provided in coded or categorised form to reduce disclosure risk.

The goal of probabilistic linking is to achieve a high link rate between the source dataset and the IDI spine, while minimising incorrect links. In the March 2025 update ('refresh') of the IDI, the link rate was 79% for MOE data and 87% for MOH data [15]. While these figures may appear low, they reflect the entire dataset as supplied by the agencies, which includes people not eligible to be in the spine (for example, short-term visitors on tourist visas or those born before administrative records contributing to the spine began). As a result, link rates for specific research populations are often substantially higher. Stats NZ measures and reports incorrect link rates (false positive rates, the proportion of incorrect links between records that do not belong to the same person), aiming to keep the rate below 2%.

In addition to domain-specific data discussed below, the IDI provides a range of core variables that are available to all IDI researchers with an approved research project. This includes demographic information such as sex/gender, birth month/year, ethnicity and region of residence, which are derived from multiple collections in the IDI using a set of specific rules.

A range of resources is available to support researchers working in the IDI. A Stats NZ Wiki, accessible within the Data Lab, provides links to data dictionaries, technical documentation, and practical tips for working efficiently with the data. Additional guidance is available from other sources, such as a guide from the Virtual Health Information Network to help new users get started in the IDI [16]. The Te Rourou IDI Search app allows users to search across available IDI datasets and variables, making it easier to locate relevant data for a project before and during analysis. The app enables keyword searching and, in some cases, provides basic metadata about variables. It is designed as a discovery tool; full data dictionaries and detailed metadata are generally only available within the secure Data Lab environment and are not publicly accessible. Researchers can also engage with the Integrated Data Commons https://idcommons.discourse.group. Launched in 2023, this secure online community brings together researchers and other IDI users to facilitate discussion, knowledge sharing, and collaboration, with the goal of supporting cross-sector research and maximising the public value of the IDI.

Administrative education data

A range of administrative data from the New Zealand education system is available in the IDI, covering enrolment, attendance, and achievement across ECE, schooling, and tertiary education. These data are supplied by the MOE, with the length of historical coverage varying by dataset. For example, tertiary education data are available from 1994, school enrolment and ECE data from 2008, so in more recent years a broader range of datasets is available, and these tend to contain more complete information [4]. Figure 2 provides an overview of important administrative education datasets, which are described in more detail below. Survey-based education datasets available in the IDI are discussed separately under the 'Census and survey data' section. All datasets

described below are person-level administrative records which can be linked over time and across domains.

New Zealand qualifications are organised within a single national framework, the New Zealand Qualifications Framework (NZQF). This framework classifies qualifications from Level 1 (basic certificates) through to Level 10 (doctoral degrees). School qualifications typically sit at Levels 1-3, while higher-level certificates, diplomas and degrees are placed at Levels 4-10 [17].

Schooling in New Zealand is compulsory from ages 6 to 16. However, most children begin school on or shortly after their fifth birthday because enrolment is permitted from age five [18]. The majority of students continue beyond the age of 16, and many remain at school past their 17^{th} birthday [19]. Secondary education typically spans Years 9 to 13 (around ages 12-18).

At the time of writing, the main secondary school qualification in New Zealand is the National Certificate of Educational Achievement (NCEA), though reforms to New Zealand's secondary school qualification framework have recently been announced. Currently, most students begin NCEA Level 1 in Year 11 and progress through Level 2 in Year 12 and Level 3 in Year 13 [20].

Education data support a wide range of research and policy analysis. For example, tertiary qualifications have been linked to school achievement and student characteristics to identify factors associated with success in higher-level education [17]. Similarly, the Tertiary Education Commission uses the IDI to monitor post-study outcomes such as earnings by field of study and education provider [18].

In recent years, New Zealand has been one of the few OECD countries without national standardised assessments for primary and secondary schooling [21]. While some voluntary assessment tools (such as e-asTTle and the Progressive Achievement Tests) are used by many schools, these data are not included in the IDI. However, the IDI does contain selected sample-based assessments, which are discussed under "Census and survey data on health and education".

Early childhood education participation

Temporal coverage: 2008 onwards.

Population coverage: Children enrolled in licensed ECE providers.

Key variables: Age at first participation, number of years the child attended, number of hours per week, provider type.

Data collection/coding: Participation usually recorded at school enrolment.

Limitations: Informal childcare not captured; relies on recall as usually collected when the child is enrolled at school; broad and imprecise measure of ECE intensity.

Early childhood education attendance

Temporal coverage: 2015 onwards.

Population coverage: Children enrolled in licensed ECE providers.

Key variables: Daily presence/absence, hours attended, provider identifier.

Limitations: Informal childcare not captured; not all ECE providers report daily attendance and reporting completeness improves over time.

School enrolment

Temporal coverage: 2008 onwards.

Population coverage: All students enrolled in primary and secondary schools.

Key variables: school identifier, enrolment dates, year level, school type (state, state-integrated, private), selected demographics.

Limitations: Administrative records may contain missing or revised enrolment dates and changes in recording practices over time.

School qualifications

Temporal coverage: 1993 onwards

Population coverage: secondary school students enrolled in programmes that lead to nationally recognised secondary school qualifications.

Key variables: Standards attempted, standards achieved, credits gained, endorsement status, and awarded qualification level.

Coding and structure: NCEA is a standards-based system. Students accumulate credits toward Level 1, Level 2 and Level 3 qualifications through internally and externally assessed standards.

Limitations: Qualification system reforms currently underway may affect comparability across cohorts. Incomplete coverage of non-NCEA school qualifications such as Cambridge International Examinations and International Baccalaureate; these qualifications are taken by a small number of students and participation is skewed towards those from higher socioeconomic backgrounds.

School leavers

Temporal coverage: 2009 onwards.

Population coverage: Students who leave secondary school in each calendar year. Also contains information on non-NZQF qualifications offered in a small number of schools, such as Cambridge International Examinations and International Baccalaureate programmes.

Key variables: Age at leaving, school last attended, highest qualification achieved, basic demographics.

Limitations: School leaver status is defined administratively and may not correspond exactly to the last day a student attended school. The dataset is finalised only after leaver status is confirmed in the following year, so there can be a delay before data become available.

School attendance

Temporal coverage: 2011 onwards.

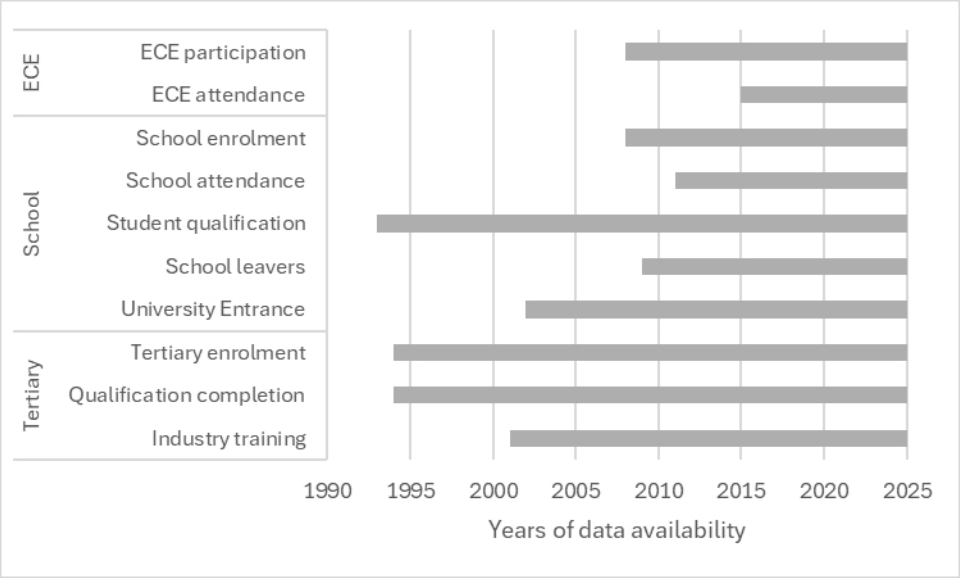
Population coverage: Students enrolled in state and state-integrated schools; limited private school data.

Key variables: Daily attendance, justified/ unjustified absence, partial absence, reason codes where recorded.

Data collection/coding: Collected via school student management systems.

Limitations: Reason codes often missing or inconsistently applied; data quality improves over time.

Figure 2: Education data in the IDI



Note: ECE = Early Childhood Education. Most education data are typically available only up to the end of the previous calendar year because key assessments are completed in November and results are processed before inclusion in the IDI.

University Entrance

Temporal coverage: 2002 onwards.

Population coverage: Records whether students have met the minimum requirement to enrol in a New Zealand university.

Key variables: Year the student attained university entrance.

Tertiary enrolments and qualifications

Temporal coverage: 1994 onwards.

Population coverage: All students enrolled at recognised tertiary providers, including universities, polytechnics, wānanga (Māori tertiary education institutions) and industry training organisations.

Key variables: Course and programme enrolments, provider identifier, field of study, enrolment dates, qualification level, and qualification completion, selected demographics.

Coding/structure: Qualifications are mapped to NZQF Levels 1-10.

Limitations: Institutional reform and funding changes may affect time trends; only New Zealand qualifications completed at recognised providers are included; in particular, overseas qualifications are not included [17].

Industry training and vocational programmes

Temporal coverage: 2001 onwards

Population coverage: All those engaged in formally recognised, workplace-based training, mainly in trades and vocational fields.

Key variables: programme type, training organisation identifier, enrolment and completion dates, credits achieved and qualifications awarded. Additional data also cover specific programmes, such as low- or foundation-level and vocationally-oriented education initiatives, adult literacy and numeracy programmes, and subsidies for employers to support the recruitment and retention of new apprentices.

Limitations: Informal workplace learning is not captured.

Administrative health data

New Zealand has a predominantly publicly funded healthcare system designed to provide universal access to health services [22]. Publicly funded health care is delivered through a network of hospitals, primary care providers, and community services. Primary care refers to first-contact, community-based health services and includes general practitioners (GPs; family doctors), who are usually the first point of contact for most health concerns, with services that are subsidised or partially subsidised. Public hospitals provide free acute and specialist care. A smaller private health sector also operates, offering faster access to non-urgent hospital and secondary specialist services for those who hold private insurance (around one in three people in 2023/4 [23]) or can pay out of pocket. The extent of private healthcare usage varies depending on the procedure; for elective procedures where public wait times are long or access is restricted there will be greater usage of private services.

Health datasets in the IDI are available at the individual level, allowing longitudinal follow-up and linkage to education, income, and other domains. The administrative health data in IDI mostly cover publicly funded services, although some privately funded hospital services are captured. There is information on hospital care, secondary specialist care, and primary care, with information being mostly complete for hospital care, less for secondary specialist, and minimal for primary care. This is driven by national reporting requirements and data collection methods, with primary care information being most limited, in part because of difficulties collating data at the national level due to variability in data collection systems between individual practices. The lack of primary

care information is a key limitation of IDI health data and the data are not well suited to research on conditions that are largely treated in primary care. Another key limitation is diagnostic information, which is recorded in some datasets (hospital admissions, mortality, cancer registrations) but is absent from others (National Non-Admitted Patient Collection NNPAC covering outpatient events, Pharmaceutical Collection, laboratory claims).

Most administrative health data are provided by Health New Zealand - Te Whatu Ora and the extent of historical data coverage varies between datasets. For example, publicly funded hospital discharges are available from 1988, mortality records from 1988, pharmaceutical dispensing from 2005, and mental health from 2000. As a result, information becomes broader and more comprehensive in later years (especially 2005 onwards). Figure 3 provides an overview of key datasets and time coverage, described in more detail below. To aid interpretation, datasets are grouped below into broad service categories reflecting common research use rather than formal system boundaries.

Health data within the IDI are widely used to answer diverse research questions across population health. Researchers use these data to explore patterns of health service utilisation, including who uses health services, how often, and in what contexts [24–29]. It also enables quantification of health outcomes, such as morbidity, mortality, and longer-term wellbeing measures, providing a population-level evidence base [30–33]. Health data are further leveraged to identify and follow cohorts of interest, for example people with a specific condition (such as autism) or those who have experienced a major health event (such as a hip fracture) [34–38]. Another important area of research is the ability to examine social determinants of health, such as education, income, employment, and housing, and how these shape patterns of health service use and outcomes across population groups [39–41]. These factors are captured through linked administrative, census, and survey data, as illustrated in Figure 1, with further detail of some datasets provided in the "Additional data in the IDI" section. Finally, the linked and longitudinal nature of the IDI allows the development of prediction models for outcomes of interest, supporting both epidemiological insight and potential policy or clinical applications [10, 31, 33, 42].

Public hospital discharges

Temporal coverage: 1988 onwards.

Population coverage: Publicly funded hospital discharge events.

Key variables: Admission and discharge dates, diagnosis and procedure codes, discharge outcomes, facility identifiers, basic demographics.

Limitations: Coding systems, variable definitions, and reporting practices have changed over time, including the transition from ICD-9 to ICD-10-AM in 2000. Short-stay emergency department events were reported inconsistently prior to 2012 [43].

Private hospital discharges

Temporal coverage: 2001 onwards.

Population coverage: Privately funded hospital discharge events, including overnight hospital stays and day-stays (including emergency stays) of greater than three hours.

Key variables: Admission and discharge dates, diagnosis and procedure codes, discharge outcomes, facility identifiers, basic demographics.

Limitations: Coverage of privately funded hospital stays is lower and more variable over time than for publicly funded admissions, with some private hospitals not submitting data. The range and completeness of clinical and cost variables are more limited and less consistent than in the publicly funded discharge data.

National Non-Admitted Patient Collection (NNPAC)

Temporal coverage: 2007 onwards.

Population coverage: NNPAC contains event-level data on publicly funded non-admitted secondary care. In practice, most records come from hospital-based outpatient clinics and emergency departments; however, the collection can also include activity delivered in other settings where it is reported through the same national system. Emergency department events that result in a hospital admission are recorded in both NNPAC and the hospital discharge dataset [44].

Key variables: Event date, facility identifier, broad service type, broad health specialty, and purchase unit (cost) codes, basic demographics.

Limitations: Includes non-admitted secondary care activity captured through the public hospital systems and does not represent the full universe of delivered outpatient care. Diagnosis information is not included.

National booking reporting system

Temporal coverage: 2003 onwards.

Population coverage: Patients who are booked or waiting for publicly funded elective medical and surgical services.

Key variables: Date of entry into the system, assessed priority, booking status, health speciality, facility identifier, exit outcome.

Limitations: Does not include privately funded elective care.

Programme for the Integration of Mental Health Data (PRIMHD)

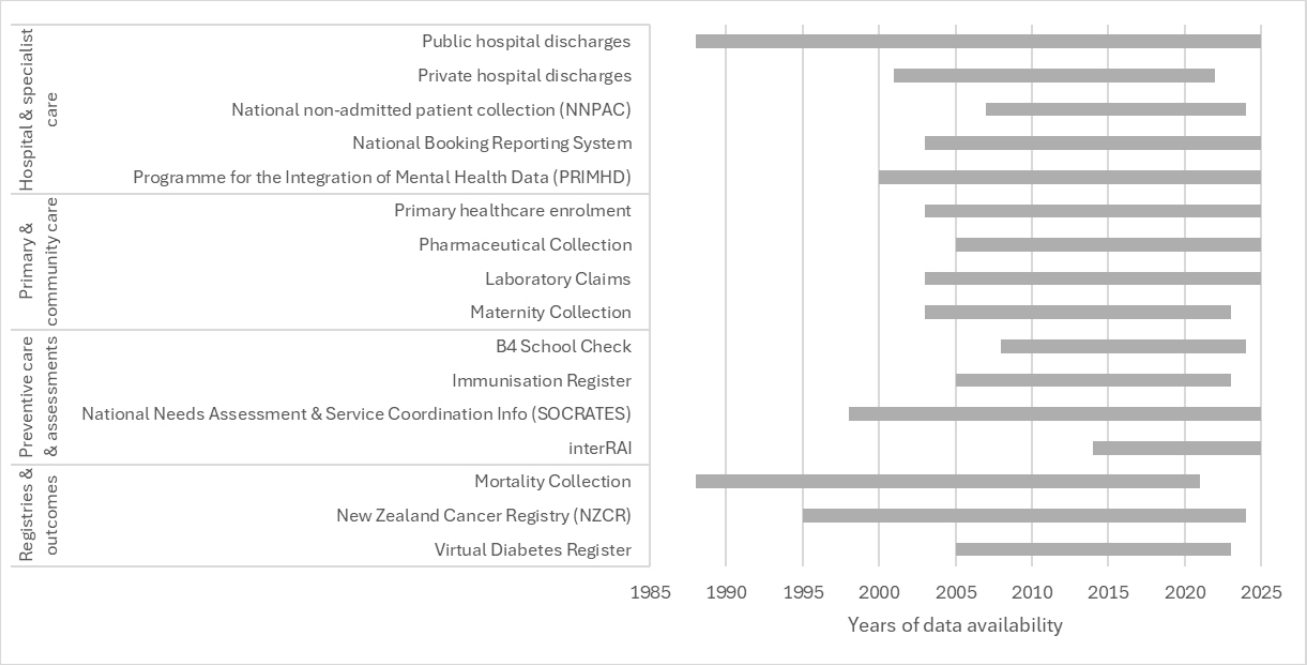
Temporal coverage: 2000 onwards.

Population coverage: People who access publicly funded secondary specialist mental health and addiction services in both inpatient and community-based care settings. These services capture approximately 3% of the population with the highest mental health and addiction needs [45].

Key variables: Contact type, team type, start and end dates of referral periods, service contact dates, servicing setting and activity type, provider and facility identifiers, and legal status under the Mental Health Act.

Limitations: Captures publicly funded secondary specialist mental health and addiction services and does not include mental health care delivered solely in primary care or privately funded settings. Diagnosis information is available in a linked

Figure 3: Health data in the IDI



Note: End-of-series truncation for some datasets reflects reporting and processing lags before data are incorporated into the IDI.

dataset using ICD-10-AM and DSM-IV codes, but coverage and completeness vary.

Primary healthcare enrolment

Temporal coverage: 2003 onwards.

Population coverage: Individuals enrolled with a Primary Health Organisation (PHO, umbrella organisations for primary care/family doctor practices).

Key variables: Enrolment status, enrolment dates, practice identifier, PHO identifier, and a quarterly indicator of contact with primary care services.

Limitations: The dataset records enrolment rather than service use and does not provide a complete record of primary care contacts. No information on diagnoses or treatments is included. Prior to 2019, enrolment information was collected through the PHO Enrolment Collection; from 2019 onwards, these data are sourced from the National Enrolment Service (NES), which may affect comparability across time [46].

Pharmaceutical Collection

Temporal coverage: 2005 onwards.

Population coverage: All dispensing of publicly funded medicines from community pharmacies.

Key variables: Chemical name and formulation, date of dispensing, cost, quantity dispensed, and days of supply.

Limitations: Dose, days of supply, and quantity dispensed have high rates of missing data. The Pharmaceutical Collection is derived from pharmacy reimbursement claims rather than clinical records, and these fields are not consistently required or captured in the claims process. This collection captures pharmaceutical dispensed but not pharmaceuticals prescribed that were never dispensed. In addition, the dataset does not record the reason or indication for prescribing and does not include over-the-counter, privately funded, or medicines administered in hospital settings. Coverage is low in the early years, and the dataset is best used from approximately 2007/08 onwards.

Laboratory claims

Temporal coverage: 2003 onwards.

Population coverage: Publicly funded community laboratory tests.

Key variables: Broad laboratory test type, service date, provider identifier.

Limitations: This collection records claims and payment information rather than clinical information; laboratory test results are not available. Privately funded tests and tests undertaken as part of hospital care are not captured. Test detail is limited to broad categories.

National Maternity Collection (MAT)

Temporal coverage: 2003 onwards.

Population coverage: People who receive publicly funded maternity care in New Zealand, covering maternity-related events from up to nine months before birth through to six weeks after birth [47].

Key variables: Maternal demographics and characteristics, pregnancy-related clinical information (including complications), labour and delivery details, birth outcomes, and infant characteristics. Data are sourced from lead maternity carers (e.g. midwives and obstetricians) and from hospital inpatient records.

Limitations: Captures maternity care delivered through publicly funded entitlements, including care provided by lead maternity carers who may also charge co-payments alongside receiving public funding (for example, private obstetricians).

Information is collected for administrative and reporting purposes, and completeness and detail may vary across care settings and over time.

B4 school check

Temporal coverage: 2008 onwards.

Population coverage: Children who receive the B4 School Check universal health and development screening assessment for four-year-olds.

Key variables: Dates of assessment; results from screening domains including vision, hearing, height/weight (growth), oral health, behavioural and developmental assessments; and indicators for referral and follow-up.

Limitations: Coverage is lower in the early years of the programme (particularly 2008–2010), and uptake is not fully universal - there are systematic differences in who receives a check (including disparities by sociodemographic factors) [48].

Aotearoa Immunisation Register (AIR)

Temporal coverage: 2005 onwards.

Population coverage: Individuals who receive immunisations that are recorded in the AIR, primarily those delivered under New Zealand's publicly funded immunisation programmes.

Key variables: Vaccine type, date of administration, selected demographics.

Limitations: The register primarily captures immunisations delivered under New Zealand's publicly funded immunisation programmes and does not fully capture all immunisation events. The set of funded immunisations recorded in the register changes over time as the National Immunisation Schedule is updated. Some immunisations received outside funded programmes or not reported to the register may be missing. Individuals may also restrict access to their immunisation records, which can affect completeness.

National Needs Assessment and Service Coordination Information (SOCRATES)

Temporal coverage: 1998 onwards.

Population coverage: Individuals who undergo needs assessments for publicly funded disability support.

Key variables: Needs assessment results, service type, service start and end dates, diagnosis information, selected demographics.

Limitations: Only includes those who engage with publicly funded Disability Support Services and therefore does not represent all people with disabilities. Eligibility criteria for Disability Support Services mean that some conditions are not supported and are not captured in SOCRATES. Eligibility rules and service definitions have changed over time, which may affect comparability. Coverage is low in the early years, and the dataset is best used from approximately 2010 onwards.

interRAI

Temporal coverage: 2014 onwards.

Population coverage: Individuals living in aged residential care or receiving publicly funded home-based support services. Completion of interRAI assessments is mandatory for entry into aged care and for people receiving publicly funded home-based support, and assessments are repeated at regular review intervals or following significant changes in health status.

Key variables: Standardised clinical assessment measures covering functional status and activities of daily living, cognitive performance, mood and mental health, comorbidities, medication use, social support, and selected demographics. Assessments are completed by trained clinicians using internationally developed, standardised interRAI instruments, enabling consistent measurement of needs across care settings.

Limitations: Coverage is limited to people assessed within aged residential care and publicly funded home-based support services and does not represent the broader population of older people or people with disabilities who do not receive these services.

Mortality collection

Temporal coverage: 1988 onwards.

Population coverage: All deaths registered in New Zealand.

Key variables: Month and year of death, underlying cause of death, and contributing causes of death, coded using ICD-9 in earlier years and ICD-10-AM in later years. Exact date of death is available within the IDI on request where required for approved research purposes (for example, defining early infant mortality).

Limitations: There can be a substantial lag in data availability, as annual mortality data are released only after cause of death coding is complete. This process can take several years for some deaths, particularly where coronial or criminal investigations are involved.

New Zealand Cancer Registry (NZCR)

Temporal coverage: 1995 onwards.

Population coverage: All registrations of primary malignant diseases diagnosed in New Zealand.

Key variables: Date of diagnosis, cancer site, ICD-10-AM diagnosis code, and healthcare location of diagnosis. Additional clinical information, including tumour grade, extent, and stage at diagnosis, is also available.

Limitations: Information on tumour grade, extent, and stage has variable completeness, with missing data in some fields, so care is required when using these variables [49].

Virtual Diabetes Register (VDR)

Temporal coverage: 2005 onwards.

Population coverage: Individuals identified as having suspected diabetes diagnoses using a composite, algorithm-based definition. Identification is based on an algorithm drawing on multiple health data sources, including hospital discharge diagnoses, outpatient contacts, laboratory tests, and pharmaceutical dispensing records [50].

Key variables: Indicator of suspected diabetes status, date of identification, selected demographics.

Limitations: As an algorithm-based construct, the VDR includes both false positives and false negatives. The identification algorithm and the quality and coverage of contributing data sources have changed over time, meaning

caution is required when using the VDR to assess trends or changes in diabetes prevalence over time [50].

Other datasets

The above datasets are those that are most likely to be of use to health researchers using the IDI. However, there are a number of other administrative health datasets from Te Whatu Ora available including:

- General Medical Subsidy claims: fee-for-service claims made by general practitioners/family doctors
- Health Tracker: health outcomes for one-year cohorts of individuals using health services
- Chronic conditions / significant health events: indicators for eight chronic conditions using health data (to 2013)
- Health Service User (HSU) population: annualised indicators of individuals using health services
- Population cohort demographics: demographics of health service users

Additionally, health information can be found in datasets from other sectors, including ACC injury claims and Ministry of Social Development Benefits data [4].

Census and survey data on health and education

In addition to the whole-population administrative data described above, the IDI also contains a range of survey datasets containing health and education information. These can be used on their own or linked to the administrative sources described above, allowing researchers to combine the strengths of different data sources. Data from the 2013, 2018, and 2023 New Zealand Censuses of Population and Dwellings provide information on a range of social determinants of health such as employment, income, education (including highest secondary school qualification and post-school qualifications such as certificates, diplomas, and degrees), disability, smoking, and housing conditions, along with demographic information on individuals, households, and families.

The New Zealand Health Survey is a cross-sectional face-to-face sample survey of adults and children that has been running continuously since 2011. Core questions include health status, service use and barriers, long-term health conditions, functional difficulties, health and risk behaviours, and weight and height measurements. Additional modules are added in some years and the content of these varies including topics such as mental health and substance use, dietary habits, racial discrimination, and migraine.

Other survey datasets in IDI that contain information on health and/or the social determinants of health include: Te Kupenga (post-census Māori social survey); the post-census Disability Survey; General Social Survey; Longitudinal Immigrant Survey of NZ (LiSNZ); Survey of Family, Income and Employment (SoFIE); Household Labour Force Survey (HLFS); and Household Economic Survey (HES).

The IDI also includes two international large-scale assessment studies of education: Programme for International Student Assessment (PISA), which measures the competencies of 15-year-old students in reading, mathematics, and science [51]; and the Programme for the International Assessment of Adult Competencies (PIAAC), which measures adult skills [52]. The 2009 and 2018 PISA cycles, as well as the 2014 PIAAC cycle, have been linked to the IDI and can be used, for example, to examine how measured skills relate to labour market outcomes captured in administrative data [8, 53].

Most survey datasets in the IDI are cross-sectional. However, some legacy surveys are longitudinal, including LiSNZ and SoFIE that followed participants over a limited number of waves. In addition, repeated Census waves (e.g. 2013, 2018, and 2023) can be linked at the individual level, enabling longitudinal analysis across collections. At the time of writing, there are also plans to add Growing Up in New Zealand, a contemporary longitudinal cohort study, to the IDI.

Additional data in the IDI

A broad range of additional administrative data sources in the IDI can be used to construct individual- or household-level characteristics, or to capture contextual factors relevant to research projects which focus on health or education [4]. These include Inland Revenue tax records on income and employment, Ministry of Social Development data on social welfare benefit receipt, police and court records on justice involvement, and border movement data from the Ministry of Business, Innovation and Employment. Household and family relationships can be inferred from birth registrations, marriage and civil union records, Census data on relationships and co-residence, and selected administrative sources such as benefit records and visa applications [54]. Employment records are also linked to the Longitudinal Business Database (LBD), a Stats NZ research database containing firm-level information from a wide range of business surveys and administrative sources [55]. This linkage enables researchers to obtain information about an individual's employer, such as industry and firm size.

These data are available at the individual level. Some contextual variables, such as geographic classifications and area-based deprivation measures, are attached to individuals based on their recorded place of residence rather than measured directly at the individual level.

The Administrative Population Census (APC) is part of Stats NZ's census transformation programme, which explores the use of administrative data to produce census-type information. It provides annual data starting 2006, and includes two derived education variables: *highest qualification* and *field of study*. These variables draw on administrative data discussed above, along with information from the 2013 Census to provide harmonised, individual level information [56].

'Code Modules' are curated tools that combine high-quality code with detailed documentation to produce standardised measures to enable working with integrated data easier and more accessible. In the education domain, two code modules are available: *educational attainment*, which creates harmonised qualification records across multiple

data sources, and *school attendance*, which standardises measures of presence and absence across schooling levels. In the health space, a module defining *Ambulatory Sensitive Care Hospitalisations (ASH)* has already been deployed, and additional modules are currently in development.

Strengths and limitations

Strengths of health and education data in the IDI include that, first, the data cover a whole-population, which means it is possible to investigate small populations that may be difficult to recruit for targeted studies. For example, studies using the IDI have investigated health and education outcomes for refugees, and for those born very pre-term [57–59]. Second, data are 'longitudinal by design' in that the data largely comprise a time-stamped record of interaction with the health and education sectors [60]. This allows explorations of antecedents and consequences of health events and education experiences. Third, use of administrative data avoids the issue of recall bias that can impact self-reports of events and experiences [61, 62]. Internationally, New Zealand is widely recognised as a leader in the development of integrated administrative data infrastructure, combining whole-population coverage, longitudinal linkage across sectors, and a mature governance and privacy framework.

Despite its strengths, there are some limitations with the IDI generally, and with health and education data in the IDI specifically. General limitations include, first, access remains restricted to secure Data Labs in New Zealand, meaning researchers must be physically present at approved sites, and international researchers wishing to use the IDI must travel to New Zealand to do so. Second, the process of checking outputs before they can be released from the Data Lab can also be time-consuming. Third, while documentation is available for datasets, including health and education datasets, metadata can sometimes be incomplete or inconsistent. In addition, some variables have non-trivial levels of missingness or variable data quality over time, reflecting differences in collection practices and the administrative purpose for which the data were originally collected.

Although linkage rates in the IDI are generally high, linkage is not complete and varies across source agencies and time periods. Incomplete or imperfect linkage may result in missing records or misclassification, and linkage quality may differ across population subgroups depending on the availability and quality of identifying information. These factors should be considered when defining analytic cohorts and interpreting results, particularly for studies focusing on small or marginalised populations.

A specific limitation of health datasets is that these are primarily service use datasets, so the data will not include health problems for which service has not been sought, and there may be biases in who receives services [63]. In addition, coverage of community-based care is incomplete; many forms of community nursing, allied health, NGO-delivered, and privately funded community services are not captured in the IDI. Further, important factors for health and education (e.g., biomarkers, aspects of the home environment) are not available in the IDI.

Conclusion

The IDI is a uniquely powerful resource for education and health research in New Zealand. By linking administrative, census, and survey data at the individual level across multiple domains, the IDI provides the foundation for analysing complex relationships and long-term outcomes with precision and scale. The depth and granularity of data enable investigations that would be difficult, if not impossible, to undertake through traditional surveys or siloed administrative sources.

This infrastructure not only supports academic research but also provides a platform for evidence-based policymaking. The IDI enables government agencies and researchers to monitor outcomes, identify trends, and target policies more effectively, drawing on data that are already collected through routine administrative processes. This allows for ongoing insights into the impacts of public services and investments without the need for costly, large-scale data collection.

Acknowledgements

We thank Charles Darr (New Zealand Council for Educational Research) and Barclay Anstiss (Ministry of Education) for their helpful comments.

Ethics statement

This article did not require ethical approval because it is a data resource profile, and no new analyses of data were undertaken.

Conflict of interests statement

None declared.

Data availability statement

No new data were created or analysed. The datasets described are part of the Integrated Data Infrastructure (IDI) managed by Stats NZ.

AI disclosure statement

The authors used AI tools for language editing and reviewed and approved all changes.

References

1. Stats NZ. Integrated Data Infrastructure. 2022; Available from: https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/.

2. Stats NZ. Estimated resident population (ERP), national population by ethnic group, age and sex, 30 June 1996, 2001, 2006, 2013, 2018, and 2023. 2025; Available from: https://explore.data.stats.govt.nz.

3. Milne, B. IDI Publications. 2025; Available from: https://www.zotero.org/groups/4681141/idi_publications/library.

4. Stats NZ. Data in the IDI. 2024; Available from: https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/data-in-the-idi/.

5. Bowden, N., et al., Associations between type 1 diabetes and educational outcomes: an Aotearoa/New Zealand nationwide birth cohort study using the Integrated Data Infrastructure. Diabetologia, 2024. 67(1): p. 62–73. https://doi.org/10.1007/s00125-023-06026-y

6. Gorman, E., et al., A national multiple baseline cohort study of mental health conditions in early adolescence and subsequent educational outcomes in New Zealand. Scientific reports, 2023. 13(1): p. 11025. https://doi.org/10.1038/s41598-023-38131-8

7. McLeod, K., et al., Sociodemographic, mental health, education, employment and income characteristics of adults with obsessive–compulsive disorder who accessed secondary health services in Aotearoa| New Zealand. Journal of the Royal Society of New Zealand, 2025. 55(6): p. 1776–1795. https://doi.org/10.1080/03036758.2024.2406827

8. Meehan, L., G. Pacheco, and T. Schober, Basic Reading and Mathematics Skills and the Labour Market Outcomes of Young People: Evidence from PISA and Linked Administrative Data. Economic Record, 2023. 99(327): p. 473–491. https://doi.org/10.1111/1475-4932.12755

9. Kokaua, J., et al., Highly qualified Māori and Pacific Peoples in Aotearoa: A study of Māori and Pacific Peoples PhD graduates (2002-2023) using administrative data in the Integrated Data Infrastructure. 2025.

10. Schluter, P.J., et al., The efficacy of preschool developmental indicators as a screen for early primary school-based literacy interventions. Child Development, 2020. 91(1): p. e59–e76. https://doi.org/10.1111/cdev.13145

11. Bowden, N., et al., Association between high-need education-based funding and school suspension rates for autistic students in New Zealand. JAMA Pediatrics, 2022. 176(7): p. 664–671. https://doi.org/10.1001/jamapediatrics.2022.1296

12. Jones, C., et al., Building on Aotearoa New Zealand's Integrated Data Infrastructure. Harvard Data Science Review, 2022. 4(2).

13. Stats NZ. How we keep integrated data safe. 2024; Available from: https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/.

14. Black, A., The IDI prototype spine's creation and coverage. 2016, Statistics New Zealand: Statistics New Zealand Working Paper No 16–03.

15. Stats NZ, Integrated Data Infrastructure (IDI) Refresh: Linking Report, Methods and Design, March 2025 Refresh. 2025.

16. Virtual Health Information Network. Getting started using the IDI. 2016; Available from: https://vhin.co.nz/guides/getting-started-using-the-idi/.

17. New Zealand Qualifications Authority. About the New Zealand Qualifications and Credentials Framework (NZQCF). 2025; Available from: https://www2.nzqa.govt.nz/qualifications-and-standards/about-new-zealand-qualifications-credentials-framework/.

18. Ministry of Education. Primary and secondary education. 2024; Available from: https://www.education.govt.nz/our-work/about-us/education-new-zealand/our-education-system/primary-and-secondary-education.

19. Ministry of Education, Retention of students in senior secondary schools, in Education Counts, Education Indicator - Education and Learning Outcomes. 2024.

20. New Zealand Qualifications Authority. How the New Zealand education system works. 2025; Available from: https://www2.nzqa.govt.nz/international/study-nz-quals/nz-education-system/.

21. OECD, Education at a Glance 2023: OECD Indicators. 2023, Paris: OECD Publishing.

22. Ministry of Health – Manatū Hauora. Health system overview and statutory framework. 2024; Available from: https://www.health.govt.nz/about-us/new-zealands-health-system/overview-and-statutory-framework.

23. Ministry of Health – Manatū Hauora. Annual Update of Key Results 2023/24: New Zealand Health Survey. 2024; Available from: https://www.health.govt.nz/publications/annual-update-of-key-results-202324-new-zealand-health-survey.

24. Carr, G., et al., Evolution of first episode psychosis diagnoses and health service use among young Māori and non-Māori—A New Zealand national cohort study. Early Intervention in Psychiatry, 2023. 17(3): p. 290–298. https://doi.org/10.1111/eip.13327

25. Charania, N.A., et al., Exploring immunisation inequities among migrant and refugee children in New Zealand. Human Vaccines & Immunotherapeutics, 2018. 14(12): p. 3026–3033. https://doi.org/10.1080/21645515.2018.1496769

26. D'Souza, S., et al., Medication dispensing for attention-deficit/hyperactivity disorder to New Zealand youth. The New Zealand Medical Journal, 2020. 133(1522): p. 84–84.

27. Figueroa, J.F., et al., International comparison of health spending and utilization among people with complex multimorbidity. Health Services Research, 2021. 56: p. 1317–1334. https://doi.org/10.1111/1475-6773.13708

28. McLay, L.K., et al., Health service utilization among autistic youth in Aotearoa New Zealand: A nationwide cross-sectional study. Autism, 2025. 29(5): p. 1143–1156. https://doi.org/10.1177/13623613241298352

29. Papanicolas, I., et al., Differences in health care spending and utilization among older frail adults in high-income countries: ICCONIC hip fracture persona. Health Services Research, 2021. 56: p. 1335–1346. https://doi.org/10.1111/1475-6773.13739

30. Cargo, T., et al., Medication dispensing among Māori and non-Māori screened for preschool ADHD. The New Zealand Medical Journal, 2022. 135(1565): p. 95–103.

31. Pugh, M., et al., Health outcomes of children in state care in Aotearoa New Zealand. Journal of Paediatrics and Child Health, 2023. 59(7): p. 895–900. https://doi.org/10.1111/jpc.16409

32. Vu, H., et al., Mortality risk among Autistic children and young people: A nationwide birth cohort study. Autism, 2024. 28(9): p. 2244–2253. https://doi.org/10.1177/13623613231224015

33. Richmond-Rakerd, L.S., et al., Longitudinal associations of mental disorders with dementia: 30-year analysis of 1.7 million New Zealand citizens. JAMA Psychiatry, 2022. 79(4): p. 333–340. https://doi.org/10.1001/jamapsychiatry.2021.4377

34. Boven, N., et al., Identifying multiple sclerosis in linked administrative health data in Aotearoa New Zealand. The New Zealand Medical Journal, 2025. 138(1612): p. 71–82. https://doi.org/10.26635/6965.6823

35. Underwood, L., et al., Long-term health conditions among household families in Aotearoa New Zealand: cross-sectional analysis of integrated Census and administrative data. The New Zealand Medical Journal, 2024. 137(1596): p. 20–34. https://doi.org/10.26635/6965.6370

36. Gibb, S., N. Brewer, and N. Bowden, Social impacts and costs of schizophrenia: a national cohort study using New Zealand linked administrative data. The New Zealand Medical Journal, 2021. 134(1537): p. 66–83.

37. Pryymachenko, Y., et al., The long-term impacts of opioid use before and after joint arthroplasty: matched cohort analysis of New Zealand linked register data. Family Practice, 2024. 41(6): p. 916–924. https://doi.org/10.1093/fampra/cmad112

38. Bowden, N., et al., Autism spectrum disorder/Takiwātanga: An Integrated Data Infrastructure-based approach to autism spectrum disorder research in New Zealand. Autism, 2020. 24(8): p. 2213–2227. https://doi.org/10.1177/1362361320939329

39. Hobbs, M., et al., The environment a young person grows up in is associated with their mental health: A nationwide geospatial study using the integrated data infrastructure, New Zealand. Social Science & Medicine, 2023. 326: p. 115893. https://doi.org/10.1016/j.socscimed.2023.115893

40. Kokaua, J., et al., Is parent education a factor in identifying autism/takiwātanga in an ethnic cohort of Pacific children in Aotearoa, New Zealand? A national cross-sectional study using linked administrative data. Autism, 2024. 28(7): p. 1667–1676. https://doi.org/10.1177/13623613231217800

41. Dawson, P., et al., Social determinants and inequitable maternal and perinatal outcomes in Aotearoa New Zealand. Women's Health, 2022. 18: p. 17455065221075913. https://doi.org/10.1177/17455065221075913

42. Mujoo, H., et al., Identifying neurodevelopmental disabilities from nationalised preschool health check. Australian & New Zealand Journal of Psychiatry, 2023. 57(8): p. 1140–1149. https://doi.org/10.1177/00048674231151606

43. Davie, G., et al. Hospital discharge data in the IDI. 2019; Available from: https://vhin.co.nz/guides/hospital-discharge-data-in-the-idi/.

44. Ministry of Health, National Non-Admitted Patients Collection (NNPAC): Data Mart - Data Dictionary. 2019.

45. Te Hiringa Mahara - Mental Health and Wellbeing Commission, Access to mental health and addiction services. 2025.

46. Health New Zealand - Te Whatu Ora. National Enrolment Service. 2025; Available from: https://www.tewhatuora.govt.nz/for-health-providers/claims-provider-payments-and-entitlements/national-enrolment-service.

47. Ministry of Health – Manatū Hauora, National Maternity Collection (MAT): Data Mart Data Dictionary. 2011.

48. Gibb, S., et al., How universal are universal preschool health checks? An observational study using routine data from New Zealand's B4 School Check. BMJ Open, 2019. 9(4). https://doi.org/10.1136/bmjopen-2018-025535

49. Gurney, J., et al., Stage at diagnosis for Māori cancer patients: disparities, similarities and data limitations. The New Zealand Medical Journal, 2020. 133(1508): p. 43–64.

50. Te Whatu Ora - Health New Zealand, Virtual Diabetes Register: Technical Guide. 2024: Wellington.

51. OECD, PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science. 2010: OECD.

52. OECD, Skills Matter: Additional Results from the Survey of Adult Skills. OECD Skills Studies. 2019: OECD Publishing, Paris.

53. Meehan, L., G. Pacheco, and T. Schober, Literacy and numeracy skills and life-course outcomes: Evidence from PIAAC and linked administrative data. Australian Journal of Labour Economics, 2024. 27(1): p. 27–72.

54. Gath, M. and C. Bycroft, The potential for linked administrative data to provide household and family information. 2018, StatsNZ.

55. Fabling, R. and L. Sanderson, A rough guide to New Zealand's longitudinal business database. 2016, The Treasury.

56. Stats NZ. Experimental administrative population census (third iteration): Information by variable. 2023; Available from: https://www.stats.govt.nz/research/experimental-administrative-population-census-third-iteration-information-by-variable.

57. Berry, M.J., et al., Gestational age, health, and educational outcomes in adolescents. Pediatrics, 2018. 142(5): p. e20181016. https://doi.org/10.1542/peds.2018-1016

58. Petrović-van der Deen, F.S., et al., Health service utilisation by quota, family-sponsored and convention refugees in their first five years in New Zealand. Australian and New Zealand Journal of Public Health, 2023. 47(3): p. 100064. https://doi.org/10.1016/j.anzjph.2023.100064

59. Marlowe, J., et al., Settlement trajectories of nearly 25,000 forced migrants in New Zealand: longitudinal insights from administrative data. Kōtuitui: New Zealand Journal of Social Sciences Online, 2024. 19(1): p. 21–44.

60. Milne, B.J., et al., Use of population-level administrative data in developmental science. Annual Review of Developmental Psychology, 2022. 4: p. 447–468. https://doi.org/10.1146/annurev-devpsych-120920-023709

61. Milne, B.J., et al., Data resource profile: the New Zealand Integrated Data Infrastructure (IDI). International Journal of Epidemiology, 2019. 48(3): p. 677–677e.

62. Jutte, D.P., L.L. Roos, and M.D. Brownell, Administrative record linkage as a tool for public health research. Annual Review of Public Health, 2011. 32(1): p. 91–108. https://doi.org/10.1146/annurev-publhealth-031210-100700

63. Svardal, C.A., et al., Prevalence of antidepressant use and unmedicated depression in pregnant New Zealand women. Australian & New Zealand Journal of Psychiatry, 2022. 56(5): p. 489–499. https://doi.org/10.1177/00048674211025699