Australian Child and Youth Wellbeing Atlas

Technical Information

Australian Child and Youth Wellbeing Atlas

TECHNICAL INFORMATION

Table of Contents

Introduction

This technical information document provides an overview of the data cleaning and data processing activities conducted in the creation of the Australian Child and Youth Wellbeing Atlas. Additionally, it also includes information about certain statistical methods employed throughout the data preparation and cleaning process.

Data cleaning

The raw data sourced from our data providers underwent a transformation process to align it with predefined guidelines and specifications. This transformation included activities like standardising variable formats, addressing inconsistencies, and managing missing data. Through these measures, we guaranteed compatibility with the Atlas database system and streamlined the integration of data.

In addition to these transformations, we focused on standardising the data to promote uniformity across variables. This entailed harmonising variable names, categorisations, and coding schemes to enhance overall data consistency.

To maintain the overall quality and integrity of the data, we identified and resolved anomalies, including outliers and data entry errors.

Geography

The "geography" field is used to indicate the geographic area to which the data pertains. This field is named using one of the following options:

ASGS 2016:

STE_CODE16

SA2_CODE16

SA3_CODE16

SA4_CODE16

LGA_CODE16

The values in this field are numeric.

For "SA2_CODE16", "SA3_CODE16", "SA4_CODE16", and "LGA_CODE16", the codes are always composed of 5 digits.

For "STE_CODE16", the codes are always composed of 1 digit. If the raw data includes character values for States, the following conversion rules should be applied: 1 = NSW, 2 = VIC, 3 = QLD, 4 = SA, 5 = WA, 6 = TAS, 7 = NT, 8 = ACT.[1]

The majority of data sets in the Atlas are at ASGS 2016.

An exception are data sets derived from the Australian Immunisation Register (ASGS 2011) and the Australian Early Development Census (ASGS 2021).

[1]

Temporal Correspondence

The Australian Bureau of Statistics (ABS) makes available an extensive range of correspondence files, in machine readable format, to help users to mathematically convert statistical data to and from geographic regions (ABS, 2021). This process, known as temporal correspondence, is used in the Australian Child and Youth Wellbeing Atlas (ACYWA) project to ensure the maintaining of a consistent experience, where possible, when viewing data in the ACYWA platform.

In Australia, most spatial data are presented using geographical resolutions which are a part of the Australian Statistical Geography Standard (ASGS), consisting primarily of statistical areas levels 1 through 4 (represented by SA1, SA2, and so on) as well as non-ABS resolutions such as the local government area (LGA) resolution. Temporal correspondence has been applied to move between editions of the ASGS, which update every 5 years alongside the release of the Australian Census. Data presented in the ACYWA platform have been collected from 2006 through to 2022 but are mainly presented using ASGS 2016 and have hence had temporal correspondence applied.

Applying Temporal Correspondence

Temporal correspondence algorithm has been developed to bring the different Australian Statistical Geography Standard (ASGS) boundaries to ASGS 2016 boundaries. This is to allow users to harmonise and analyse data from different time periods and geographies consistently.

There are two types of temporal correspondence: forward correspondence and backward correspondence.

Forward correspondence

This is mainly correspondence of ASGS 2006 or ASGS 2011 to ASGS 2016. Forward correspondence is explained using simple graphic illustrations.

Let's assume A and B are geographical areas from ASGS 2011 and C, D and E are areas from ASGS 2016. The requirement is to convert the count values at ASGS 2011 to ASGS 2016.

Figure 1 represents the relationship of the geographical areas. As can be seen in Figure 1, *A* and *B* donated their areas to create C, D and E. Area D has been created using both areas A and B. The ratio on each rows shows how much is donated by each donator *where* $a + b = 1$ *and* $c + d = 1$.

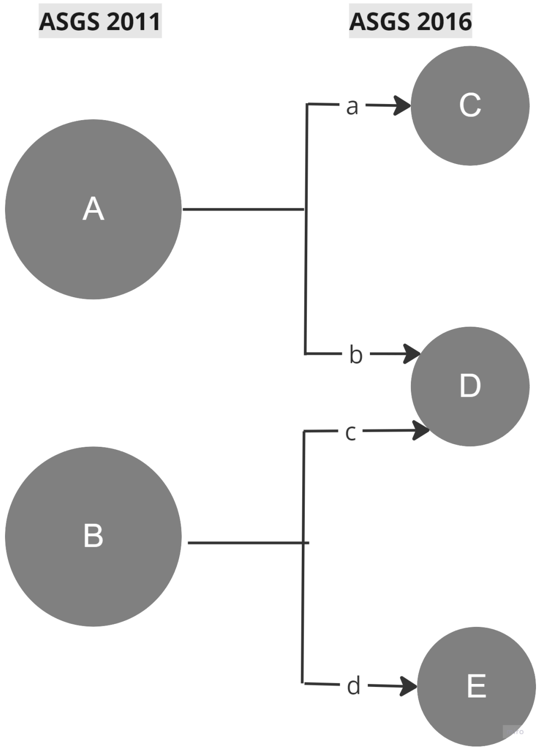

Figure 1: Forward correspondence

The count values of C, D and E from the above figure are calculated as follows, assuming all areas represent count data.

- $C = a * A$
- $D = b * A + c * B$
- $E = d * B$

Backward correspondence

This section mainly concerns the correspondence of ASGS 2021 backwards to ASGS 2016.

In Figure 2, we have A, B and F as geographical areas from ASGS 2016 and C, D, E, G and H as areas from ASGS 2021 all representing count data requiring conversion from ASGS 2021 to ASGS 2016.

Figure 2 represents the relationship of the geographical areas. As can be seen in Figure 2, areas A and B donated their counts to create C, D and E. Area D has been created using both areas A and B. The ratios shown between each areas show how much has been donated by each area *where* $a + b = 1$, $c + d = 1$, *and* $g + h = 1$.

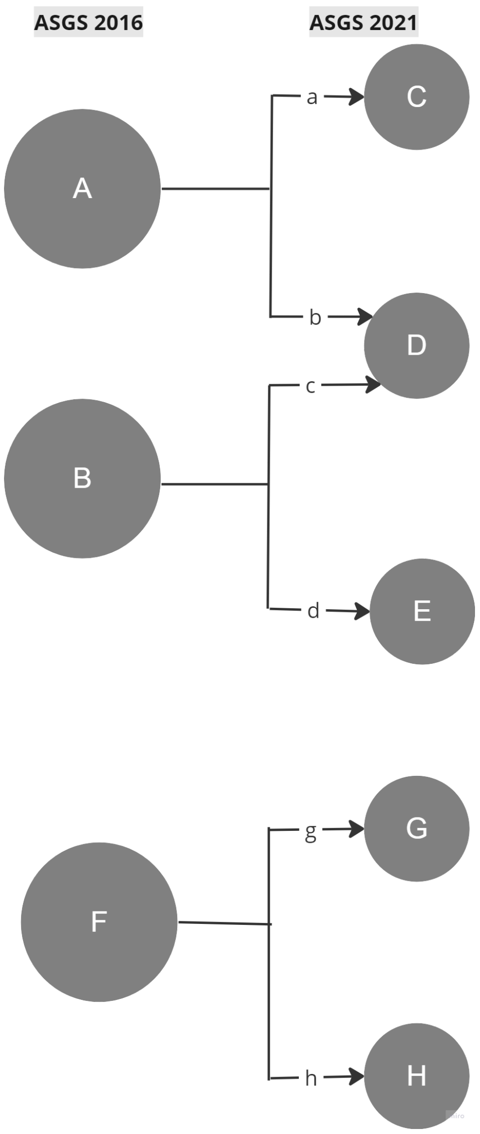

Figure 2: Forward Correspondence

For example, the count values of A, B and F are calculated as follows.

- $F = G + B$
- If *b < 0.1*, $A = C$ (as the ratio is considered as an insignificant contribution)
- If *c < 0.1*, $B = E$ (as the ratio is considered as an insignificant contribution)
- If *b > 0.1 and c > 0.1*, the values of areas A and B will be suppressed, as the full information is not available to recover what proportion of D should donated back to A and B, respectively.

In real geographical areas there is a very small number of cases requiring suppression (i.e., where both $b \text{ and } c < 0.1$. Therefore, the loss of information during backwards temporal correspondence is low.

Obtaining Correspondence Tables

The ABS have developed the following geographical correspondence files used in the ACYWA project. See below a brief list of concordance file links for reference.

2006 to 2016 (SLA - SA2; SSD - SA3; SD - SA4)

- https://data.gov.au/dataset/ds-dga-c9c33a59-f716-4c5a-b519-56ab178fc50b/details?q=2006%20correspondence

2011 to 2016 SA2, SA3, SA4

- https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument
- https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016?OpenDocument

2016 to 2021 SA2, SA3, SA4, LGA

- https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/correspondences

Notably, the ABS have provided uncertainty columns alongside correspondence ratios which indicate when the correspondence process is of “Good”, “Acceptable”, or “Poor” quality, based on the population and size of both regions involved in the correspondence process.

When the result of the correspondence process is “Poor”, values have been suppressed.

Sex

The "sex" column is used to indicate the sex of the individuals in the dataset. This column includes categories for male and female. The “all” column is a summation of male and female counts.

Where "all" counts encompass more than just male and female categories, this information is specified in the metadata for the respective indicator or data collection.

Not all indicators include disaggregation by sex.

Age group

The "age_group" field is used to indicate the age range of the individuals in the dataset. This column includes categories in the format of a hyphen between two numbers (e.g. "18-24" or “0-4”).

If the raw data was presented with school year, the following conversion table was used to populate the "age_group" column:

Table 1 - conversion between school year and age groups

School year	Corresponding age_group
Prep/Pre-Primary	5-5
Year 1	6-6
Year 2	7-7
Year 3	8-8
Year 4	9-9
Year 5	10-10
Year 6	11-11
Year 7	12-12
Year 8	13-13
Year 9	14-14
Year 10	15-15
Year 11	16-16
Year 12	17-17

Time frame

There are two options for indicating the time frame of the collected data:

1. The "calendar_year" field is in numeric format with the values following the pattern of four digits (e.g. "2022").

2. The "year_range" field is in numeric format with the values following the pattern of four digits separated by a dash (e.g. "2010-2011"). For this option, the Atlas will show the latter year of the year range ie. “2011” for "2010-2011".

Uncertainty

This field indicates the level of uncertainty present in the raw data. A value of 0 represents no uncertainty, while a value of 1 or 2 represents a level of uncertainty indicated by one or two asterisks respectively (* or **). These uncertainty levels are provided by several data custodians to indicate that data may not accurately represent the true population in cases where estimation is used or where data have been deliberately and randomly altered to increase data privacy. Where uncertainty is produced during the temporal correspondence process (detailed above), the level of uncertainty is added with existing uncertainty included in the raw data.

Data with a high level of uncertainty (**) has been suppressed.

Missing values

When missing data was encountered in the cleaning process, an "NA" was entered to indicate that the data was not available.

Cell Suppression

Cell suppression involves obscuring or removing specific cells in a dataset that may contain personally identifiable information. This technique is crucial when working with datasets that require careful handling of privacy concerns.

To implement cell suppression, we adopt a standardized approach in this project. Count and proportion values below a certain threshold are replaced with the value **9999999**, indicating that the data has been suppressed. The specific rules for cell suppression depend on the level of data privacy required.

As a general guideline, we suppress count or proportion values when the number of observations (n) is between 1 and 4. This threshold ensures that individual identification is prevented. Additionally, if high uncertainty is detected in the data, we also apply cell suppression to the corresponding count or proportion values.

Cell suppression is performed during the final stage of the data preparation process. After the indicators have undergone thorough review and the data format has been validated, the reviewer applies the cell suppression function to the dataset. This step ensures that sensitive information is adequately protected while maintaining the integrity and quality of the data.

Data Cleaning Tools and Software

R was selected as the primary tool for data cleaning in this project due to its extensive capabilities in statistical computing and data manipulation. R provides a range of specialised packages and libraries, such as dplyr and tidyr, designed for data cleaning. These packages offer functions for data reorganisation and transformation. With its diverse ecosystem, R is well-suited for addressing various data cleaning challenges, including the handling of missing values, outlier detection and treatment, data integration, and aggregation.

To facilitate the data cleaning process and promote collaboration, the Atlas Data Cleaning repository was established on GitHub. This repository functions as a centralised hub for all the R scripts used in the data cleaning phase. Each indicator being cleaned has its own dedicated script, organised and named according to the corresponding indicator code. This structured approach allows for easy navigation and access to the specific scripts required for cleaning individual indicators. By utilising the Atlas Data Cleaning repository, team members can efficiently share, review, and update the R scripts, promoting consistency and reproducibility across the project.

The R scripts make use of various functions to reshape the data into tables suitable for importing into the Atlas database and to aggregate and summarize the data as necessary.

Validation and Quality Control

For data validation and quality control, we use R Markdown. R Markdown streamlines our validation and quality control processes, automating formatting checks and outlier detection. Custom functions and scripts within R Markdown help validate data formatting, ensuring adherence to guidelines.

R Markdown generates comprehensive R Markdown Document (RMD) files for systematic validation. These RMD files contain predefined rules for checks, including acceptable values for columns, proper naming formats, and regex pattern validations for various fields.

Statistical Methods

Crude rates

All indicators included in the Child and Youth Wellbeing Atlas are at the population level, therefore the assumption is that everyone is at risk for the whole of the year (or years) of interest, as opposed to person-time at risk. Therefore, an event rate is calculated by dividing the total number of new cases of an event in a specified period (usually one year) by the average number of people in the population during the same period. This is then multiplied by 1,000, 10,000 or 100,000 and presented as a rate per 1,000, 10,000 or 100,000 people per year respectively.

Depending on the data source, population denominators are calculated as the average of the size of the population at the start and at the end of the period of interest or estimated from Census data. These basic rates are called 'crude' rates because they describe the overall incidence in a population without taking any other features of the population into account (e.g., age structures).

Age-specific rates

A crude comparison may have little meaning if the groups that are being compared have very different age structures. To address this, we can calculate rates for individual age groups, known as age-specific rates. The rate in a particular age group can then be compared between geographic areas. This process can be extended to calculate individual rates for other groups, for instance male and female (sex-specific rates), and for different racial or socioeconomic groups.

Standardised rates

Comparisons between rates may become difficult if age-specific rates are presented for a large number of different age-groups. An alternative is to summarise or combine these age-specific rates using the process of direct standardisation. This involves calculating the overall incidence or mortality rate that would be expected in a 'standard' population (i.e. population with a hypothetical age structure) if it had the same age-specific rates as the study population. Direct standardisation requires:

1. The age-specific event rates in the study population and
2. The age distribution of the standard population

When the population being studied are not known but the total number of events is known, then the indirect standardisation is commonly used. The indirect method is also often used for small populations where fluctuations in age-specific rates can affect the reliability of rates calculated using the direct method. There are many similarities as well as differences between the two methods. However, the two methods will yield comparable results in most cases. It could be argued that the choice of a standard population is more important than the choice of the direct or indirect method. The standard population used in the Atlas for purposes of indicator comparisons is the State population.

Where the data allows, the direct method of age-standardisation is the method chosen for use in the because of its advantages over the indirect method when comparing Aboriginal and non-Aboriginal mortality rates, disease incidence and prevalence rates over time. Also note that indirect standardisation fixes the quantity of interest (i.e. age-specific rates) as the standard, and then compares the effect of differences in age-structure in two or more populations. For this reason, it is less useful as a public health comparator than direct standardisation.

Formula[1]:

Direct method

$SR= (SUM\ (r_i * P_i))/SUM\ P_i$

Indirect method

$SR=(C/SUM(R_i *p_i))*R$

Where:

SR is the age-standardised rate for the population being studied
r_i is the age-group specific rate for age group i in the population being studied
P_i is the population of age group i in the standard population
C is the observed number of events* in the population being studied
SUM($R_i p_i$) is the expected number of events in the population being studied
R_i is the age-group specific rate for age group i in the standard population
p_i is the population for age group i in the population being studied
R is the crude rate in the standard population

* 'Events' can include deaths, incident or prevalent cases of disease or other conditions, or health care utilisation occurrences.

Prior moving averages

A prior moving average (PMA) is defined as the average of the span of series values preceding the current value. The span is the number of preceding series values used to compute the average.

PMA =

$$\frac{M_{t-2} + M_{t-1} + M_t}{P_{t-2} + P_{t-1} + P_t}$$

A 3-year prior moving average combines a sequence of 3 years of data prior to, and including, the selected year. Similarly, a 5 year prior moving average combines a sequence of 5 years of data prior to, and including, the selected year.

Series based on prior moving averages are presented as overlapping sequences until the most recent year is included. Moving averages make it possible to combine more years of data to maximize sample size at each point while maintaining data confidentiality.

Age groupings

As there is little difference in the resulting rate ratios and rate differences using five- or ten-year age-groupings, we typically follow the usual convention of using five year age-groupings in the calculation of directly age-standardised rates. However, if the distribution of the data across age-groups requires collapsing of age-groups to overcome small numbers, then 10-year age-groupings may be used.

Also, due to little or no difference in rate differences produced using 0-4 compared to using <1 and 1-4 age groups in the estimation of age-standardised rates, we follow the usual practice of using the 0-4 age group as the youngest age group in the calculation of age-standardised rates. This only applies to the calculation of age-standardised rates, and does not preclude presenting age-specific rates and distribution of events (e.g. deaths) for <1 and 1-4 age groups). If these age groups are not used, the actual age groups are detailed in notes accompanying the age standardised population rate information. Standardised rates are generally multiplied by 1,000 or 100,000 to avoid small decimal fractions. They are then called standardised rates per 1,000 or 100,000 population.

Standardised ratios

The indirect method is also used to calculate standardised mortality ratios (SMRs) and other standardised ratios, for example for health service utilisation and other events. These ratios express the overall experience of a comparison population in terms of the standard population by calculating the ratio of observed to expected deaths in the comparison population. This is calculated by dividing the observed number of deaths by the expected number. Sometimes the SMR is multiplied by 100 to express the ratio as a percentage, although this is not universally accepted. Not multiplying by 100 has the benefit of being able to say that the SMR was, for example, 2.3 times that expected rather than 130% higher.

Association and Cause

All data presented in the Child and Youth Wellbeing Atlas are not based on specific information about individuals but relate to the number of events (or deaths) in a population relative to the size of that population (often an estimate from the ABS census).

When comparing the strength of the relation between indicators, caution should therefore be made when trying to relate the occurrence of an event to potential causes. For example, there may be a statistical association between two indicators in the Child and Youth Wellbeing Atlas, which may lead to an assumption of a real association. However, there should be consideration of other possibilities that may be the cause of such associations, such as chance, bias or confounding. Three important 'alternative explanations' for associations are:

1. chance (random error)
2. bias (systematic error) and
3. confounding

Chance or random error

Random error is the divergence, by chance alone, of a measurement from the true value. There are three main sources of random error: biological variation (natural variation of measurement depending on an individual's biology), measurement error (imprecision inherent in the measuring system being used), and sampling error (selection of sample from whole population).

It is impossible to completely remove random error that has resulted from chance. Therefore, when examining an association between two indicators in the CDA, it is important to consider how likely it is to be a real effect, or whether it could have arisen by chance. Whilst associations between indicators should not be ignored, any interpretations of these relationships taken on its own should be cautious, and acknowledgement should be made of the possibility that it could just reflect the effect of chance.

Small numbers

The number of cases and/or deaths from a specific type of indicator for a given sex, age-group and year can be very small. Due to a large degree of year-to-year statistical fluctuation in these small numbers, great care should be taken when assessing apparent differences involving small numbers and measures based on small numbers.

Bias

Many potential sources of bias have been identified in epidemiological studies, but all fall into two main areas: bias with respect to who gets into the study (selection bias) and bias with respect to the information we collect from, or on, these people about their exposures and their diseases (measurement, information or observation bias). Bias, also known as systematic error, is potentially more problematic than random error because it's much harder to know what effect it might have on an outcome. The most common systematic errors with administrative data involve underreporting of activity for a specific population, inaccurate re-coding of spatial information, or differences in data entry protocols (Ardal & Ennis, 2001).

Confounding

Confounding is where an apparent relationship between an exposure and an outcome is due, in whole or in part, to a third factor that is associated with both the exposure and the outcome of interest. Confounding is a mixing of effects because the effect of the exposure we might be interested in is mixed up with the effect of some other factor. Age, sex and socioeconomic status (SES) are common confounders.

References:

1. Australian Institute of Health and Welfare (AIHW). Age-standardised rates. AIHW (METeOR). Available from: http://meteor.aihw.gov.au/content/index.phtml/itemId/327276

2. Australian Institute of Health and Welfare (AIHW). National Healthcare Agreement: PI 07–Infant and young child mortality rate, 2017, AIHW (METeOR). Available from: _http://meteor.aihw.gov.au/content/index.phtml/itemId/630004

3. Australian Institute of Health and Welfare (AIHW) 2011, Principles on the use of direct age-standardisation in administrative data collections: For measuring the gap between Indigenous and non-Indigenous Australians. Available from: https://www.aihw.gov.au/reports/indigenous-australians/principles-on-the-use-of-direct-age-standardisatio/contents/table-of-contents

4. Ardal S, & Ennis S 2001, Data detectives: Uncovering systematic errors in administrative databases. In Proceedings: Symposium 2001, Achieving Data Quality in a Statistical Agency: A Methodological Perspective.

5. Australian Bureau of Statistics (ABS) 2013, Statistical Language. Available from: http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language

6. Australian Bureau of Statistics (ABS) 2021, Correspondences. Available from: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/correspondences