

## RESEARCH ARTICLE

# Complex early childhood experiences: Characteristics of Northern Territory children across health, education and child protection data

Lucinda Roper<sup>1,2\*</sup>, Vincent Yaofeng He<sup>2</sup>, Oscar Perez-Concha<sup>1</sup>, Steven Guthridge<sup>2</sup>

**1** Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia, **2** Centre for Child Development and Education, Menzies School of Health Research, Charles Darwin University, Darwin, Australia

\* [roper.lucinda@gmail.com](mailto:roper.lucinda@gmail.com)



## OPEN ACCESS

**Citation:** Roper L, He VY, Perez-Concha O, Guthridge S (2023) Complex early childhood experiences: Characteristics of Northern Territory children across health, education and child protection data. PLoS ONE 18(1): e0280648. <https://doi.org/10.1371/journal.pone.0280648>

**Editor:** Ralph C. A. Rippe, Leiden University: Universiteit Leiden, NETHERLANDS

**Received:** April 27, 2022

**Accepted:** January 4, 2023

**Published:** January 19, 2023

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0280648>

**Copyright:** © 2023 Roper et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The study datasets contain sensitive personal information and are held on a secure cloud-based server with restricted access. Access requires the approval of the ethics

## Abstract

Early identification of vulnerable children to protect them from harm and support them in achieving their long-term potential is a community priority. This is particularly important in the Northern Territory (NT) of Australia, where Aboriginal children are about 40% of all children, and for whom the trauma and disadvantage experienced by Aboriginal Australians has ongoing intergenerational impacts. Given that shared social determinants influence child outcomes across the domains of health, education and welfare, there is growing interest in collaborative interventions that simultaneously respond to outcomes in all domains. There is increasing recognition that many children receive services from multiple NT government agencies, however there is limited understanding of the pattern and scale of overlap of these services. In this paper, NT health, education, child protection and perinatal datasets have been linked for the first time. The records of 8,267 children born in the NT in 2006–2009 were analysed using a person-centred analytic approach. Unsupervised machine learning techniques were used to discover clusters of NT children who experience different patterns of risk. Modelling revealed four or five distinct clusters including a cluster of children who are predominantly ill and experience some neglect, a cluster who predominantly experience abuse and a cluster who predominantly experience neglect. These three, high risk clusters all have low school attendance and together comprise 10–15% of the population. There is a large group of thriving children, with low health needs, high school attendance and low CPS contact. Finally, an unexpected cluster is a modestly sized group of non-attendees, mostly Aboriginal children, who have low school attendance but are otherwise thriving. The high risk groups experience vulnerability in all three domains of health, education and child protection, supporting the need for a flexible, rather than strictly differentiated response. Interagency cooperation would be valuable to provide a suitably collective and coordinated response for the most vulnerable children.

committee and data custodians. For applications for data access, please contact the Menzies Data-linkage Program Leader at [steve.guthridge@menzies.edu.au](mailto:steve.guthridge@menzies.edu.au), or the Menzies Data Management Team at [datamanagement@menzies.edu.au](mailto:datamanagement@menzies.edu.au).

**Funding:** The study was completed by LR as a component of her post-graduate studies. The Child and Youth Development Research Partnership (CYDRP) data repository is supported by a grant from the NT Government. The funders of the repository had no role in study design, data preparation, analysis, decision to publish or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Background

As a society we aim to support the optimal development of children. Social and economic factors, such as parental education, income, drug and alcohol use or housing shape the environments that children are exposed to, which may support or hinder their development [1]. These environments effect development in all domains, including physical, mental and emotional health as well as educational attainment. The same biopsychosocial factors which can detriment a child's health can also cause unstable and potentially unsafe home environments, which can result in family contact with the child protection system. For instance, poverty and crowding are associated with higher levels of respiratory and ear infections and are also associated with contact with the child protection service [2, 3]. Poor physical health and contact with child protection services are both associated with reduced educational attainment [4–6]. Early life experiences have a long term effect on later health and function, as key social, emotional, cognitive and physical capabilities are learnt at this stage [7–9]. Early life intervention (generally accepted as the first 8 years of life) is therefore a recognised priority area. This can include intervening when societal expectations are not met—for instance in response to low school attendance, child maltreatment or poor health.

## Data from different agencies to support the needs of vulnerable children

Australia, like all developed countries, has multiple government and non-government organisations that support children by identifying families at need and providing services. Specifically, this occurs across the domains of health, education and child protection. Data from these agencies is not only useful to record services for administrative reasons, but can enable research to inform responses, particularly when data can be shared and linked across agencies [10]. This is because there is increasing recognition that interagency collaboration is essential to holistically support these children, given the shared dimensions of health that influence outcomes across these areas [11–13]. Linked, population-wide, administrative health, education and social service data are increasingly recognised as powerful tools to provide insight into factors affecting childhood development, to improve services and interventions [10].

## Administrative datasets available in Australia for study of child development

Child protection data typically includes observational data of children who have contact with the child protection service and has been used to predict a range of developmental outcomes, including physiological outcomes, emotional functioning, cognitive and academic performance and also healthcare service utilisation [14–21]. More recently, population level, administrative child protection data has been used to build predictive models to identify children at highest risk of harm—and offer those families early intervention before the predicted harm has occurred [22]. There are however ethical concerns surrounding the use of predictive models in social services, particularly relating to privacy, stigma and 'self-fulfilling prophecy' for families identified as high risk [23].

Health data used in the study of vulnerable children typically measures the impact of disease on early childhood development [24–26], via data on hospital admissions in the early years or using perinatal data to extract birth risks. Chronic conditions have been the most extensively studied, and a diagnosis of any chronic condition is associated with poorer educational outcomes at a range of ages [27–31], with comparable findings in studies of specific conditions such as asthma [25, 30], cerebral palsy [26] or epilepsy [24]. Similar effects have been found for

childhood infections [32, 33], injuries [34–36], and overall number of ED presentations and total hospital inpatient days between birth and age seven [5].

Given the impact of childhood illnesses on development, an alternative paradigm is consideration of the preventability of these conditions—many illnesses are associated with poverty and failure to receive preventative care or timely care (e.g. vaccination, dental care, asthma management). The Australian Institute of Health and Welfare (AIHW) list of ambulatory care sensitive conditions (ACSC) has been used for pediatric research [37], but misses numerous common pediatric conditions, such as gastroenteritis [38]. Several pediatric-specific indicators of avoidable hospitalization have been developed in the United States, United Kingdom and New Zealand (NZ) [39, 40]. The NZ indicator suite includes admissions that may be avoidable via policy measures that influence the socioeconomic determinants of health [41]. This allows a broader range of conditions to be captured and prevents unrealistic expectations of the role of healthcare in prevention, when larger social policies affecting socioeconomic gradients also play a role. The NZ indicator suite has been used in Australia to compare avoidable hospitalizations in New South Wales Aboriginal and non-Aboriginal children [41, 42]. The pediatric indicator found more avoidable hospitalizations than the adult indicator, and these hospitalizations were more associated with social and health disadvantage at birth [38].

The Australian education system collects data on enrolment and attendance, and also performance on nationally standardised tests at years 3, 5, 7 and 9 (National Assessment Program—Literacy and Numeracy, NAPLAN) [43]. More recently, Australia has also begun collecting information through the Australian Early Development Census (AEDC) [44]—a three-yearly nationwide data collection of early childhood development, completed when children are in their first year of full-time school. Education data is used by government to monitor performance against national targets, and by educators to assess the causal impact of a wide range of factors on performance (i.e. pre-school attendance, teacher qualifications, school funding model) [45, 46]. Furthermore, this data has been used by human development economists to understand how schooling relates to later outcomes—e.g. correlations with school dropout, delinquency, and involvement with the legal system [47, 48]. Australian studies indicate that early variation in attendance rates persist over the school career [49]. Whilst there has traditionally been an assumption that attendance and performance are linked, it is not clear how policies aimed at improving attendance affect performance, particularly for Aboriginal schoolchildren [50].

### Limitations of current studies

Administrative Health, Education and Child Protection Services (CPS) data have rarely been linked together in Australia; datasets are typically considered either individually or in pairs. For instance, linked hospital and CPS data have been analysed in New South Wales (NSW) and Western Australia (WA), and it has been established that maltreated children have higher rates of healthcare utilisation [51–53]. Research using these data sources together has focused on improving identification of children at risk of harm [51, 54] or defining healthcare costs to build an economic case for policy change [17, 52, 55–57]. Linked hospital and education data have also been studied together, to clearly establish that illness has a detrimental effect on educational attainment [24, 25, 33]. CPS and educational data have been studied together, in the NT and elsewhere [58], with many studies focusing on the effect of specific elements of child abuse on educational outcomes. International studies indicate that adverse outcomes such as school dropout, are related to whether multiple types of abuse occurred and the chronicity and severity of maltreatment [14, 15, 18–21, 59–62].

Not only has previous consideration of these datasets been limited to pairs; frequently one dataset is assumed to represent the outcome—for instance hospitalisation as an outcome of child maltreatment. However, increased hospital utilisation and CPS involvement may sometimes have a reverse relationship: an ill child may cause household stress and increased propensity to maltreatment, as well as increased exposure to health professionals, who are the main CPS reporter group in early childhood [63]. Therefore, although there has been research into vulnerable children in one or two dimensions, their overlapping risks across these three domains are not fully understood. There may be subpopulations with different patterns of risk, who may benefit from differentiated responses.

### Context of the current study

The Northern Territory (NT) of Australia is the smallest of Australia's states and territories by population, comprising approximately 1% of the total Australian population [64]. The NT is characterised by its large Aboriginal population, comprising approximately 30% of the NT population, compared to the national average of 3% of the population [65]. The NT Aboriginal population has a relatively young age structure, with a median age of 26 years compared to 35 years for the non-Aboriginal population [64] and Aboriginal children make up 43% of all NT children [66].

Aboriginal Australians experience disadvantage in almost all measures of health and welfare, and in the NT, the Aboriginal population experience disproportionate levels of poverty, crowded housing and poor health [67–70]. Although Aboriginal children make up 43% of NT children, they comprise 82% of children in contact with the CPS [71] and also have higher rates of avoidable hospitalisations [42, 72] and lower rates of school attendance [49, 73]. School attendance, particularly in the early years, allows children to develop a foundation for later education and learning. NT Aboriginal students have average attendance rates below 60% [74], whereas the Australian Curriculum, Assessment and Reporting Authority uses 90% attendance rate as a key performance measure to assess adequate attendance level [75].

In Australia, the suite of government research and policies aimed at reducing inequality between Aboriginal and non-Aboriginal Australians are known as “Closing the Gap” [76]. Initially referring to closing the life expectancy gap, this term also applies to gaps in privilege and attainment in a wide variety of areas—including health, education and employment. By 2018, the 2014 target: to close the gap between Aboriginal and non-Aboriginal students' attendance rates, within 5 years, had not been met [77]. The 2020 updated National Agreement on Closing the Gap contains 17 national targets, three of which are specifically relevant to this project:

- Aboriginal and Torres Strait Islander children are not overrepresented in the child protection system
- Aboriginal and Torres Strait Islander children thrive in their early year, and
- Aboriginal and Torres Strait Islander children are engaged in high quality, culturally appropriate early childhood education in their early years

In the NT, government agencies involved in the identification of vulnerable children include the NT Department of Health (Health), NT Department of Education (Education) and the Department of Territory Families, Housing and Communities (child protection service (CPS)). Each has their own data systems and policies for sharing and management of data. There is currently limited understanding of how children are treated by or notified to multiple services. However, there is increasing recognition that children may be seen by all three NT government agencies and that a multi-agency person-centred response could be beneficial

[78–80]. Thus, the NT Government has collaborated with the Menzies School of Health Research's Centre for Child Development and Education (CCDE) and invested in the Child and Youth Development Research Partnership (CYDRP). This collaborative research partnership between CCDE and the NT Government supports the ongoing maintenance and development of an extensive data repository to be used for approved research projects that inform the health and wellbeing of NT children

The motivation for this present study was to discover if routinely collected data on NT children can be used to separate children into groups who may benefit from differentiated responses, including collaborative inter-agency interventions for the most vulnerable. To remain sensitive to the complexity of the relationships between different indicators of disadvantage it was considered that an exploratory, data-driven approach, such as clustering, was most likely to return useful information, irrespective of causality, to inform planning for multi-agency approaches.

## Methods

### Data sources

This project utilises four datasets from the Northern Territory (NT)—the Perinatal Data Register, Hospital Admissions, School Enrolment and Attendance and Child Protection Service (CPS) contacts. The data are held in a data repository containing de-identified, unit-level linked records for NT children across a total of 14 datasets. The repository is longitudinal data-set, developed through the Child and Youth Development Research Partnership (CYDRP) between Menzies School of Health Research and multiple NT Government agencies, including the departments of Health, Education, and Territory Families, Housing and Communities (child protection) [81]. Initial linkage was conducted by SA NT Datalink [82], using probabilistic methods to match the records for children across multiple datasets, with clerical review of uncertain matches. This process is confirmed to result in 99.6% accuracy for completed links [82]. SA NT DataLink creates a unique linkage key for each child and provides this to data custodians. Each data custodian then creates a de-identified research dataset containing only the linkage keys and approved research variables, which is provided to researchers. Researchers are then able to merge records for the same child across multiple datasets. A full description is available elsewhere [83].

The NT Perinatal Data Register (perinatal data) is a statutory collection containing demographic, antenatal and birth information for all births in the NT and was used to define the cohort. The Hospital Admissions dataset (hospital data) contains information on all admissions to public hospitals in the NT. The Education Enrolment and Attendance dataset (education data) contains information on daily school attendance for children attending NT Government schools. Approximately 70% of all NT children attended government schools in 2020. Nationally, 65% of all Australian children and 83% of Aboriginal children attended government schools [84]. The CPS dataset contains information on all notifications of suspected maltreatment, with further details for those notifications that are investigated and substantiated.

### Study population

The study population comprised all children born in the NT between 2006–2009 and present in the NT Government school attendance records in Year 1 (usually age 5–6). Of the 15,284 children born during the selection period, 8,267 children were retained in the study after excluding those with no records in NT Government school attendance records in Year 1. Children born to mothers who were resident of the major centres of Darwin and Alice Springs and their hinterlands at time of birth were classified as urban residents; all other children were classified as remote residents.

## Research question

Can we discover distinct groups of NT children, at the end of their first year of formal schooling, with varying patterns of contact with the CPS, Health and Education services?

## Analysis approach

### 1. Initial descriptive analysis

Initial descriptive analysis was undertaken, including counts and percentages for categorical variables and mean, median and standard deviation for continuous variables, and data visualisations including histograms and bar plots, to provide an overview of the data. The results were stratified by Aboriginal status, as disparities in the health, social and educational outcomes between Aboriginal and non-Aboriginal children in the Northern Territory were expected.

## Health data

Three components of the hospital data were explored:

- Any hospitalisation (binary) from birth to age 5: overall and with a specific diagnosis. The specific diagnoses explored were ‘avoidable hospitalisations’—defined as a diagnosis on the Anderson avoidable hospitalisation list [56]. This is a list of 26 conditions represented by 130 ICD-10-AM codes, which can be found in the appendix of the original paper [85]. The birth admission was excluded.
- Number of hospitalisations (count) from birth to age 5: overall or with a specific diagnosis (diagnosis codes defined as per above).
- Length of stay per hospitalisation (average per child)

## Child protection service data

Five child protection variables were explored. For each variable, both binary (any experience of outcome) and count (number) were explored. In the NT there is a statutory requirement that requires all adults to report any incident, to either the Department of Territory Families Housing and Communities, or the police, if they have a reasonable belief that a child has been or is likely to be harmed [86, 87]. [S1 Appendix](#) contains further detail on rationale for inclusion of each child protection variable into the models.

- Notifications: a notification is a report made to the CPS by a reporter that a child is experiencing or at risk of neglect or abuse.
- Substantiations: after a notification, some notifications are investigated, based on a risk assessment. Approximately one third of investigated notifications result in substantiations (i.e. confirmed abuse or neglect).
- Abuse type: physical, emotional, sexual abuse or neglect. A description of physical, emotional, sexual abuse or neglect can be found here [87]. Emotional harm occurs when a child’s social, emotional or cognitive development is impaired or is at significant risk as a result of their parents’ or carers’ persistent failure to meet the child’s emotional need for love and security, or their psychological needs for stimulation and nurturing. This includes exposure to family or domestic violence, which requires an automatic report for emotional abuse.

- Reporter type: this is the recorded occupation of the reporter. This has been grouped into eight categories: community members (including family or self-reporting), child protection officer, school personnel, police, health professional, non-government organisation, other professional (including social workers) and not stated.
- Substantiation descriptor: if a notification is substantiated, then an additional categorical variable—the substantiation descriptor—is recorded. Although there are many different substantiation descriptors, alcohol and other drugs (AOD) or domestic violence (DV) are the most common and are explored here.

## Educational data

The educational engagement measure chosen was Year 1 school attendance, defined as total days attended in Year 1, divided by total days expected to attend. Expected to attend is a variable provided by school data to represent ideal attendance for that specific child, i.e. a child who moved schools and was only enrolled for 6 months, is only expected to attend for 6 months

### 2. Data-driven approach: clustering

In our context, a cluster is defined as a collection of children aggregated together because of similarities in regard to the variables used in this study, using a distance-based similarity measure, such as Euclidian Distance. We used a data-driven approach, in which representations of distinct groups of NT children are automatically learned using unsupervised machine learning techniques. Unsupervised learning, such as clustering, uses data that is unlabelled, meaning that children have not been pre-assigned any labels or categories by the researcher. Instead of using a pre-defined outcome to form groups, the algorithm discovers patterns in the data using only the input data [88]. The k-means clustering methodology was used.

It is common practice to stratify analysis by Aboriginal status, because of the major differences in a range of conditions between Aboriginal and non-Aboriginal populations. In this study, however, we did not split the cohort by Aboriginal status for the clustering, for two reasons. Firstly, keeping both groups allowed assessment of the plausibility of the clusters. Given known differences in health, education, and social outcomes between Aboriginal and non-Aboriginal children, we would expect most non-Aboriginal children to be in a relatively healthy cluster. Secondly, because there are no requirements for the clusters to be the same size, a small, highly vulnerable cluster can be discernible, irrespective of the number of low-risk children in the cohort. Unlike traditional modelling which may struggle to estimate parameters for both groups, clustering is very flexible and can accommodate diverse samples.

### Clustering methodology: K-means

K-means is a simple and efficient algorithm, popular in health research [89], a full explanation can be found in [S2 Appendix](#). In summary, it is a distance-based algorithm (Euclidian Distance), which follows the below steps:

- a. Pre-select the number of clusters
- b. Initiation step: several cluster centres are randomly chosen
- c. Assignment/Expectation step: the Euclidean distance between each datapoint and each initial cluster centre is calculated and each datapoint is assigned to the cluster it is nearest to
- d. Update/Maximisation step: cluster centres are updated to be the mean of all points that were assigned to that cluster

- e. The algorithm continues to move between steps c and d until convergence is reached (when the assignment of points to a cluster no longer changes)

The 'ideal' number of clusters will result in the most compact and separated clusters. We used the silhouette coefficient to assess this, which is explained in the supplementary material. It is a score that captures dissimilarity between clusters and similarity within clusters [90].

## Dimensionality reduction

K-means cannot deal with the 'curse of dimensionality', meaning that its performance can suffer in high-dimensionality datasets. With increasing dimensions, each observation in the dataset appears similarly distant to all others as the more dimensions involved, the greater chance that a difference apparent in one dimension becomes nullified by similarity in another [91, 92]. A more detailed explanation of this is in [S3 Appendix](#).

There has therefore been substantial work done on variable subset selection for use in clustering [93–97]. It is a challenging problem, firstly, because there are no labels to allow for evaluation of variable importance based on classification accuracy (supervised machine learning) or univariate relationship with the outcome. Secondly, because the number of clusters isn't predetermined, and the cluster number affects variable importance, it is difficult to unpick the entwined issues of cluster number and variable selection [96].

We have taken the following approach:

- a. Dimensionality reduction, via three approaches outlined below
- b. Clustering on reduced dimensionality data
- c. Map the reduced dimensionality data, with their cluster assignments, to the full dataset, to characterise the clusters

We therefore created three separate cluster structures, by using three separate approaches to dimensionality reduction. These are explained in detail in the supplementary material. Firstly, we used two well-established pre-clustering variable selection methods: filtering and feature extraction via principal component analysis. Secondly, we combined these with a more novel approach, which iteratively uses post-clustering variable importance rankings, to define a new variable subset.

- Method 1: Expert selection: the use of expert knowledge, guided by the principle of diversity, to choose a subset of variables which represent unique aspects of the data whilst having a presumed relationship to the outcome of interest. This is also known as a filter method [94, 97].
- Method 2: PCA: PCA is a method of automatic dimensionality reduction that projects the original data into a smaller number of dimensions [98]. PCA was carried out using the scikit-learn package, set to reduce the data to 4 dimensions [90].
- Method 3: post-clustering variable extraction via decision trees, following the steps below:
  - a. create clusters based on the variable subset resulting from method 1 (features extracted by human expert) and method 2 (features extracted by PCA)
  - b. Predict cluster membership using the Extra Tree Classifier in scikit-learn [90] and rank all original 43 variables based on their contribution to predicting cluster membership—for details on this variable importance metric, see [99]
  - c. Select the top 10 ranked variables from each model and use PCA to reduce these to four dimensions

## Post hoc analysis—mapping clusters done in reduced dimension to original data

Cluster labels were assigned to each data point, and a descriptive analysis was performed, within clusters, for variables within the health, CPS, education and perinatal datasets, with additional demographic information relevant to the NT. Informative names were then assigned to each cluster based on patterns revealed in the descriptive analysis.

Given that each cluster was generated in a different, reduced-dimensionality subspace, there is no objective method to assess which method generated the 'best' clusters, as we cannot compare their silhouette scores. We have therefore been informed by the concept of evidence accumulation [100]. This was originally a method to combine multiple cluster structures within a process known as consensus clustering, essentially treating each cluster label as a 'vote' then reassigning final cluster membership based on total votes across all cluster structures (for more detailed explanation please see references) [100, 101]. The term has also been used to describe the human-machine interaction required to qualitatively assess if the underlying concepts communicated in different clustering structures are consistent with each other [100].

### 3. Software and machine learning model parameters

Analysis was undertaken in Stata version 15 (Stata Corporation, College Station, TX, USA) Python and R [102, 103]. The scikit-learn package in Python [90] was used for the PCA and clustering algorithms and the Keras package [104] was used to rank feature importance.

### 4. Resources and data governance

Data access was subject to the conditions for use of Child and Youth Development Research Partnership (CYDRP) data repository. All project data was stored and accessed from the secure CYDRP data server in keeping with the CYDRP data security declaration and the conditions of research ethics approval in which the student is a named investigator (HREC 2016–2708). This project was approved by the CYDRP Steering Committee and reviewed by the CYDRP First Nations Advisory Group who approved the methodology, aims and objectives of the study.

## Results

### Initial descriptive analysis

Table 1 contains descriptive statistics of our study cohort ( $n = 8,267$ ). There was a distinct difference in remoteness between Aboriginal and non-Aboriginal children. The majority of Aboriginal children were from remote regions (their mothers reported place of residence at the time of their birth) while the majority of non-Aboriginal children were from urban regions. There were higher rates of young maternal age, premature birth, maternal alcohol use and smoking in pregnancy in Aboriginal compared to non-Aboriginal children. There were higher rates of hospitalisation and child protection contact amongst Aboriginal children than non-Aboriginal children (Table 1). In terms of school experience, Aboriginal children experienced higher school mobility and lower school attendance (Table 1).

### Clustering

Three variable sets were used to create three sets of clusters, as described in the methods. Table 2 below summarises the models, including the number of clusters, size and descriptive names for the clusters. For the list of the features used in methods 1 to 3 please see S4 Appendix.

**Table 1. Characteristics (%) of our study cohort born in the NT from 2006 to 2009.**

Characteristics	Aboriginal (n = 4,624)	non-Aboriginal (n = 3,643)
<b>Sex</b>		
Male	48.0	49.6
Female	52.0	50.4
<b>Remoteness</b>		
Urban	35.0	77.7
Remote	65.0	22.3
<b>Perinatal factors</b>		
Maternal anemia in pregnancy	4.9	0.9
Gestational diabetes	10.8	6.9
Maternal STI in pregnancy	9.0	1.0
<7 antenatal care visits attended	37.4	16.0
Mother <18 years old at delivery	11.1	1.2
First born child	30.6	41.2
Preterm birth	12.7	6.9
Low birth weight	12.3	6.0
APGAR<7	2.6	1.3
Neonatal resuscitation required	54.0	52.0
Maternal alcohol use in pregnancy*	13.3	6.3
Maternal smoking in pregnancy**	50.7	18.9
<b>Hospital diagnosis from birth to age 5</b>		
Avoidable hospitalisation	59.7	19.9
Infancy related condition	32.6	18.1
Gastrointestinal disease	29.0	5.1
Skin infection	21.3	2.0
Nutritional deficiency	20.7	0.7
Acute bronchiolitis	18.6	3.4
Injury	17.0	8.0
Bacterial pneumonia	16.1	2.7
Otitis media	16.0	3.1
Acute URTI^ excluding croup	12.9	5.0
Dental condition	7.1	1.7
<b>Child Protection contact from birth to age 5</b>		
Any notification	46.9	12.3
Any substantiation	23.0	2.9
Notification by police	25.4	5.4
Notification by health professional	21.2	2.2
Notification by school personnel	4.3	1.9
Substantiated attributed to DV	13.2	1.8
Substantiation attributed to AOD	9.0	0.9
<b>Year 1 school experience</b>		
Attending more than one school	29.2	4.5
School attendance rate (mean)	71	93

\*744 missing (Aboriginal children), 236 missing (non-Aboriginal children)

\*\*856 missing (Aboriginal children), 386 missing (non-Aboriginal children)

^Upper Respiratory Tract Infection

<https://doi.org/10.1371/journal.pone.0280648.t001>

**Table 2. Result for three clustering methods.**

Cluster method	Number of clusters, K	Cluster names	Proportion of cohort in each cluster (%)
Model 1: Expert selection	5	Non-attenders	20
		Thriving	65
		Neglect	4
		Abuse	7
		Ill	4
Model 2: PCA	4	Mixed, low vulnerability	14
		Thriving	79
		Mixed, high vulnerability	4
		Ill	4
Model 3: Post cluster variable extraction	5	Non-attenders	12
		Abuse	3
		Thriving	77
		Neglect	2
		Ill	7

<https://doi.org/10.1371/journal.pone.0280648.t002>

### Post-hoc analysis of clusters

As evident in Table 2, each of the three methods separated the dataset into subpopulations which demonstrated that the most vulnerable clusters experience overlapping social, health and education risks. All methods estimated that the most vulnerable children comprise 10–15% of the population and separated these children from the low-risk clusters. The post-hoc analysis was performed using the 17 key variables identified in method 3 of dimensionality reduction, as these were the most informative and interpretable of the original 43 variables. We present abbreviated results of Model 1 below, with full post-hoc analysis of Model 1, Model 2 and Model 3, found in S5 Appendix.

The five clusters identified in Model 1 were named *neglect*, *abuse*, *ill*, *thriving* and *non-attenders*. A descriptive summary, across the four clustering variables and four key demographic characteristics, is presented in Table 3. Two of these factors were included as they are significant for the NT context—the proportion of each cluster that was Aboriginal and the

**Table 3. Post-hoc cluster analysis of Model 1.**

Cluster	Non-attender	Thriving	Neglect	Abuse	Ill
<b>Cluster size</b>					
number of children in each cluster (n)	1639	5361	343	594	330
proportion of each group in study cohort (%)	20	65	4	7	4
<b>Median</b>					
Year 1 attendance rate	46	92	74	84	60
Hospitalisations by 5	2	0	3	2	9
Neglect notifications by 5	0	0	4	1	1
Abuse notifications by 5	0	0	1	3	0
<b>Proportion (%)</b>					
Aboriginal	97	36	91	79	94
Remote	68	18	41	26	62
Born premature	13	8	16	12	26
Alcohol in pregnancy	10	7	33	18	19

<https://doi.org/10.1371/journal.pone.0280648.t003>

proportion from a remote area (meaning that their mothers lived in a remote location at time of birth). Two perinatal factors were also included—maternal alcohol use in pregnancy and prematurity, as these factors have known links to childhood ill health and neglect.

The '*neglect*' group contained 343 (4%) of children. In this group, the median number of notifications for neglect was 4, median school attendance was 74%, compared to the national average of 93% for Year 1 school attendance in 2018 [105]. Children in this group had high hospital admissions, with median of 3. Although primarily experiencing neglect, there was also evidence of risk of abuse, with a median number of 1 abuse notification.

The '*abuse*' group contained 594 (7%) children. Median notifications for abuse were 3, with median school attendance of 84%—higher than the neglect group. This group had a median of 2 hospitalisations. The abuse group had some risk of co-existing neglect, with median of 1 neglect notification by age five.

The '*ill*' group consisted of 330 (4%) children, with a median of 9 hospital admissions. This group had a median of 1 neglect and 0 abuse notifications. This group had the lowest school attendance—a median of 60%.

The '*thriving*' group was the largest group, with 5361 (0.65) children. This group had a median of 0 abuse or neglect notifications, and a median of 0 hospital admissions. The median school attendance was 91.45%.

Finally, the '*non-attending*' cluster had low school attendance, of 46%. The non-attending cluster had a moderate number of hospitalisations, with a median of 2 and a median of 0 neglect or abuse notifications.

The '*non-attending*', '*neglect*' and '*ill*' clusters were predominantly comprised of Aboriginal children (91–97%), and a substantial proportion were from remote areas (41–68%). The '*ill*' group had the highest proportion of premature birth (26%) and the '*neglect*' group had the highest proportion of maternal alcohol use in pregnancy (33%).

## Discussion

### Pattern of vulnerability across the domains of education, health and child protection

Clustering NT children based on their health, education and child protection data consistently identified a group of highly vulnerable children who comprise up to 15% of the cohort, and experience overlapping risks of poor health, low school attendance and contact with the child protection system. Models 1 and 3 identified *neglect*, *abuse*, *ill*, *thriving* and *non-attenders* clusters, whereas Model 2 identified *ill*, *thriving*, *mixed low-risk* and *mixed high-risk* clusters, rather than separating them out by *type* of vulnerability.

The segregation of vulnerable children by risk type in Model 1 and Model 3 is considered to be more explanatory of the data and is consistent with previous CYDRP research—specifically the finding that, in the NT, a subset of children only receive reports for neglect or only receive reports for emotional abuse, while a small minority receive overlapping reports across all types of abuse—presumably representing different subpopulations [81]. The implication of this finding is that, depending on the type of abuse notification, children may come from different risk clusters and require different interventions. Whilst out-of-home care is avoided where possible it is sometimes necessary to protect children from immediate harm from violent and dangerous homes and removal may be required for some of the 7% of children identified as belonging to the *abuse* cluster [106]. On the other hand, children experiencing neglect may benefit from support through family and parenting interventions, a more suitable response for the 4% identified in the *neglect* clusters by this study [107].

The *ill* group had the highest rate of premature birth, which is a known risk factor for a variety of later childhood illnesses, particularly lung disease [108, 109]. This may have contributed to the ill health of some members of this cluster, however, the *neglect* and *ill* groups particularly overlapped. The 4% of children in the *ill* cluster tended to have some contact with CPS via neglect notifications, and the 4% of children in the *neglect* cluster also tended to have higher than average hospitalisations. As discussed earlier, this likely represents a two-way relationship with children experiencing neglect being more likely to become unwell—possibly due to poorer hygiene, diet, reduced preventative health and reduced parental supervision (injury) [70]. Conversely, non-neglected children who have frequent hospital visits have higher exposure to health professionals, a primary reporter source for neglect notifications [81]. Of the highly vulnerable clusters, the *neglect* and *ill* clusters have lower median school attendance than the abuse cluster. School absenteeism has been found to be associated with any contact with the child protection system, however with a particular association with neglect substantiations [4]. This has been hypothesised as a direct effect of reduced parental supervision [4]. The *neglect* cluster also had the highest proportion of maternal alcohol use in pregnancy. This may be for two reasons. Firstly, risky alcohol use in pregnancy triggers an automatic mandatory report of neglect to the CPS at the time of birth. Secondly, previous research has found that children of mothers with alcohol use disorders are at higher risk of later contact with the CPS, particularly with neglect [110, 111].

As discussed below, these clusters with overlapping risks of ill health and neglect are primarily from remote areas (unlike the abuse cluster which is primarily urban) and are 94% Aboriginal. The remote families these clusters represent are potentially a target for evidence based remote parenting support programs that incorporate cultural elements of early child rearing significant to Aboriginal people [112, 113]. There is some evidence that a shortage of services through remote health clinics could contribute to increased hospitalisation [114] but not to the extent to fully explain this cluster. Rather than simply increasing health services, the combined risks of neglect and ill health may be better targeted via parenting interventions.

Whilst the cluster algorithms generally separate children with high rates of hospitalisations, abuse or neglect, they do overlap. This provides support for the concept of holistic, ‘all-of-child’ programs, in the early years of life when the pattern of neglect or abuse is not clear. In these high-risk groups, the increased hospitalisation and notifications are present from the first few months of life, and could trigger a holistic, rather than a single system target intervention (i.e. just health or just CPS). Nurse home visits are one example of a holistic intervention and reduce later involvement with the child protection system [115], and have been implemented in parts of the NT [116]. Given these three clusters *combined* comprise less than 15% of the cohort, it may be feasible to aim for progressive universalism, with this group specifically targeted by effective, resource intensive interventions such as nurse home visits.

The highly vulnerable groups differed geographically depending on type of risk—the *ill* group was majority remote, the *neglect* group nearly half remote, whereas the *abuse* group was 74% urban. As discussed previously, exposure to domestic violence is considered a form of emotional abuse and leads to a mandatory report from police, if they attend an incident [86]. An analysis of >80,000 cases of intimate partner violence in the NT (2009–2014) found higher incidence in urban centres [117]. In remote areas, there were an average of 202 incidents per 1000 population, compared to average of 782 per 1000 population in urban areas [117]. Furthermore, a retrospective analysis of data specific to Royal Darwin Hospital (the sole trauma referral centre in the NT), from the Australia New Zealand Trauma Registry found that in the NT, being injured from intimate partner violence in an urban or remote, as opposed to very remote, location carried higher odds of previous presentations with intimate partner violence [118].

## School attendance rate and vulnerability

To interpret the significance of school attendance in defining the clusters, we have used the concept of evidence accumulation [100] to extract the most significant and consistent insights. Firstly, all three models identified a clearly thriving group, with low hospitalizations and CPS contact. Notably, for Model 2 and Model 3, some children in the *thriving* cluster had low school attendance, suggesting that children can be otherwise thriving, but with poor school attendance. In Model 1, where no children with poor attendance existed in the *thriving* cluster, the *non-attenders* cluster closely mirrors the thriving group in terms of low CPS and health risks, but has lower attendance.

Given these thriving non-attenders lack the risks traditionally associated with poor attendance (neglect, abuse or poor health), poor attendance may be more a reflection on the suboptimal availability of western school options. There has traditionally been a deficit framing surrounding the education of Aboriginal children growing up in remote communities [119]. The fact that some children with poor attendance are thriving in other ways suggests that they are not in deficit, but are perhaps learning a different set of skills and values from those traditionally taught and measured in western education [120]. A similar comment was made after the 2015 introduction of Direct Instruction as a new curriculum for remote NT schools. This has been described as a prescriptive teaching methodology, originally designed to bring developmentally delayed American children up to a minimum standard on a specific set of skills, and mis-applied to remote NT schools, who have majority Aboriginal, multilingual and not developmentally delayed students. It was suggested that poor attendance might be a consequence of a program that bored students and disenchanted teachers [121]. The contentious recommendations to close remote high schools [122] are similarly concerning, driven by a desire for statistical parity among Aboriginal and non-Aboriginal students, but disregarding the documented emotional toll caused by dislocation of students from their families and communities, and its association with past policies of forced assimilation [123, 124]. The NT Government funding of schools based on attendance was likely to further exacerbate these issues, as poorly attended schools that most need support are then least able to access it—and without quality education options available at homelands/outstations, many students dis-engage from schools [125]. Fortunately, NT Government funding of schools is under review and is proposed to shift to enrolment based funding [126].

The *non-attenders* cluster may benefit from education programs which inspire them to engage, and from attendance programs that are strengths-based and community embedded. An example of valuing Aboriginal cultural knowledge in education is the 'two way' curricula, which embed Aboriginal and Western knowledge into the formal curriculum—i.e. as promoted by the 'Growing Our Own' program of teacher education [127, 128]. Strengths-based attendance programs include the Clontarf Foundation, which aims to improve education, self-esteem and life skills for young Aboriginal and Torres Strait Islander men, and programs such as the Stronger Smarter Sisters program, implemented at Katherine High School [129, 130]. Finally, the Remote School Attendance Strategy is a program running in 83 schools across Australia, which employs community members to help improve children's school attendance. The nation-wide impact of the Remote School Attendance Strategy appears to be modest but has been embraced by some communities [131–133].

## Limitations

Health data was collected for administrative purposes and therefore does not provide granularity or textual notes to fully explain admissions. Furthermore, health data is restricted to hospital admissions and does not capture other health service events including clinical assessments

in remote health clinics, a key component of health care in the remote NT. Secondly, our key education variable was attendance in Year 1, which is a coarser measure of developmental vulnerability than a purpose designed measurement, such as the AEDC. Future work may investigate how the 'non-attending but thriving' cluster scored on the AEDC. Thirdly, notifications or substantiations recorded in the CPS data are only a proxy for abuse and neglect of a child and may not capture every case. Also, our study was likely to underestimate the domestic violence rates as only domestic violence substantiated notifications were included in the study due to data limitations.

The completeness of reporting of events is also an important factor when considering the *non-attenders* cluster, as data collection may be limited by service availability if the regions with poorer school attendance also have a more general limit on services including fewer potential CPS reporters. It has been suggested that intimate partner violence, for instance, may be under-reported in very remote settings [118]. This is less likely in the case of child maltreatment, firstly given NT mandatory reporting laws require every person in the NT to report child abuse or neglect, and secondly given the transparency provided by communal living in remote communities, it is likely that most maltreated children will be recorded, even if not every event [86].

A further limitation is the absence of methods to establish the best fit of the three cluster structures. There are no objective methods to determine the most important or informative of our clustering results. Therefore, there may be other ways to interpret the patterns existing in the data, which are not shown in our exemplary model [94]. To account for this limitation our analysis drew on the concept of evidence accumulation [100]. In this study, each of the clustering methodologies provided evidence that there was a minority group of children who experienced complex, overlapping risks and were differentiated primarily based on maltreatment type and hospitalisation level. The incidentally noted intermediate risk group, the *non-attenders* were present in only some cluster methodologies, and should therefore be interpreted with more caution.

This project provides evidence to support the development of cross agency, collaborative interventions in early childhood. This analysis secondly identified that poor school attendance, particularly in remote areas, may be independent of other early childhood risks, which may support the concept of providing interventions informed by the pre-existing strengths of children in remote communities. Whilst the majority of the discussion has a focus on Aboriginal children, a substantial minority (up to 20%) of the highest risk groups are non-Aboriginal children who also require effective early childhood interventions.

## Conclusion

The use of unsupervised machine learning techniques on a large, linked health, education and child protection dataset, of 8267 children from the Northern Territory of Australia, produces clusters that describe differing patterns of risk. Results from each of the three clustering methods found that 10–15% of children are very high risk, and this is differentiated into children who are predominantly ill, experience neglect or experience abuse, and all high risk groups have low school attendance. These highest risk groups experience vulnerability across all three domains, supporting the need for early, holistic intervention before more specific approaches may become necessary as these groups differentiate later in childhood. Interagency cooperation is central to delivering a suitably collective and coordinated response for the most vulnerable children. A secondary finding was of a large group of, predominantly Aboriginal, children who have poor school attendance in their early years but are otherwise thriving and may be a target for strengths-based remote school attendance programs.

## Supporting information

**S1 Appendix. Rationale for inclusion of CPS variables into models [14, 18, 20, 58, 134].**  
(DOCX)

**S2 Appendix. Details on description of K-means [89, 135].**  
(DOCX)

**S3 Appendix. Detail on dimensionality reduction methods [91, 93–97].**  
(DOCX)

**S4 Appendix. Detail on the 3 dimensionality reduction methods [58, 90, 93, 94, 98, 99, 136–140].**  
(DOCX)

**S5 Appendix. Full model results.**  
(DOCX)

## Author Contributions

**Conceptualization:** Steven Guthridge.

**Data curation:** Vincent Yaofeng He, Steven Guthridge.

**Formal analysis:** Lucinda Roper, Vincent Yaofeng He.

**Funding acquisition:** Steven Guthridge.

**Investigation:** Lucinda Roper, Steven Guthridge.

**Methodology:** Lucinda Roper, Vincent Yaofeng He, Oscar Perez-Concha.

**Project administration:** Steven Guthridge.

**Software:** Lucinda Roper, Vincent Yaofeng He.

**Supervision:** Steven Guthridge.

**Visualization:** Lucinda Roper.

**Writing – original draft:** Lucinda Roper.

**Writing – review & editing:** Lucinda Roper, Vincent Yaofeng He, Oscar Perez-Concha, Steven Guthridge.

## References

1. Moore TG, McDonald M, Carlon L, O'Rourke K. Early childhood development and the social determinants of health inequities. *Health Promotion International*. 2015; 30(suppl\_2):ii102–ii15. <https://doi.org/10.1093/heapro/dav031> PMID: 26420806
2. Smith-Vaughan H, Leach A, Shelby-James T, Kemp K, Kemp D, Mathews J. Carriage of multiple ribotypes of non-encapsulated *Haemophilus influenzae* in aboriginal infants with otitis media. *Epidemiology and Infection*. 1996; 116(2):177–83. <https://doi.org/10.1017/s0950268800052419> PMID: 8620909
3. Coulton CJ, Crampton DS, Irwin M, Spilsbury JC, Korbin JE. How neighborhoods influence child maltreatment: A review of the literature and alternative pathways. *Child Abuse and Neglect*. 2007; 31(11–12):1117–42. <https://doi.org/10.1016/j.chiabu.2007.03.023> PMID: 18023868
4. Armfield JM, Gnanamanickam E, Nguyen HT, Doidge JC, Brown DS, Preen DB, et al. School absenteeism associated with child protection system involvement, maltreatment type, and time in out-of-home care. *Child Maltreatment*. 2020; 25(4):433–45. <https://doi.org/10.1177/1077559520907682> PMID: 32166980

5. Evans A, Dunstan F, Fone DL, Bandyopadhyay A, Schofield B, Demmler JC, et al. The role of health and social factors in education outcome: a record-linked electronic birth cohort analysis. *Plos one*. 2019; 14(8):e0220771. <https://doi.org/10.1371/journal.pone.0220771> PMID: 31398202
6. Considine G, Zappalà G. The influence of social and economic disadvantage in the academic performance of school students in Australia. *Journal of Sociology*. 2002; 38(2):129–48.
7. Shonkoff JP. Leveraging the biology of adversity to address the roots of disparities in health and development. *Proceedings of the National Academy of Sciences*. 2012; 109(Supplement 2):17302–7. <https://doi.org/10.1073/pnas.1121259109> PMID: 23045654
8. Currie J, Rossin-Slater M. Early-life origins of life-cycle well-being: Research and policy implications. *Journal of Policy Analysis and Management*. 2015; 34(1):208–42. <https://doi.org/10.1002/pam.21805> PMID: 25558491
9. Gilbert R, Widom CS, Browne K, Fergusson D, Webb E, Janson S. Burden and consequences of child maltreatment in high-income countries. *The Lancet*. 2009; 373(9657):68–81. [https://doi.org/10.1016/S0140-6736\(08\)61706-7](https://doi.org/10.1016/S0140-6736(08)61706-7) PMID: 19056114
10. Brownell MD, Jutte DP. Administrative data linkage as a tool for child maltreatment research. *Child Abuse and Neglect*. 2013; 37(2–3):120–4. <https://doi.org/10.1016/j.chiabu.2012.09.013> PMID: 23260116
11. Putnam-Hornstein E, Webster D, Needell B, Magruder J. A public health approach to child maltreatment surveillance: Evidence from a data linkage project in the United States. *Child Abuse Review*. 2011; 20(4):256–73.
12. O'Donnell M. Towards prevention—a population health approach to child abuse and neglect: health indicators and the identification of antecedent causal pathways. [Doctor of Philosophy—Unpublished]. In press 2009.
13. Australia OS. Information sharing guidelines for promoting safety and wellbeing. 2013. 2020.
14. English DJ, Upadhyaya MP, Litrownik AJ, Marshall JM, Runyan DK, Graham JC, et al. Maltreatment's wake: The relationship of maltreatment dimensions to child outcomes. *Child Abuse and Neglect*. 2005; 29(5):597–619. <https://doi.org/10.1016/j.chiabu.2004.12.008> PMID: 15970327
15. Jackson Y, McGuire A, Tunno AM, Makanui PK. A reasonably large review of operationalization in child maltreatment research: Assessment approaches and sources of information in youth samples. *Child Abuse and Neglect*. 2019; 87:5–17. <https://doi.org/10.1016/j.chiabu.2018.09.016> PMID: 30392993
16. Chartier MJ, Walker JR, Naimark B. Separate and cumulative effects of adverse childhood experiences in predicting adult health and health care utilization. *Child Abuse and Neglect*. 2010; 34(6):454–64. <https://doi.org/10.1016/j.chiabu.2009.09.020> PMID: 20409586
17. Bonomi AE, Anderson ML, Rivara FP, Cannon EA, Fishman PA, Carrell D, et al. Health care utilization and costs associated with childhood abuse. *J Gen Intern Med*. 2008; 23(3):294–9. <https://doi.org/10.1007/s11606-008-0516-1> PMID: 18204885
18. English DJ, Bangdiwala SI, Runyan DK. The dimensions of maltreatment: introduction. *Child Abuse and Neglect*. 2005; 29(5):441–60. <https://doi.org/10.1016/j.chiabu.2003.09.023> PMID: 15970319
19. Lau AS, Leeb RT, English D, Graham JC, Briggs EC, Brody KE, et al. What's in a name? A comparison of methods for classifying predominant type of maltreatment. *Child Abuse and Neglect*. 2005; 29(5):533–51. <https://doi.org/10.1016/j.chiabu.2003.05.005> PMID: 15970324
20. English DJ, Graham JC, Litrownik AJ, Everson M, Bangdiwala SI. Defining maltreatment chronicity: Are there differences in child outcomes? *Child Abuse and Neglect*. 2005; 29(5):575–95. <https://doi.org/10.1016/j.chiabu.2004.08.009> PMID: 15970326
21. Litrownik AJ, Lau A, English DJ, Briggs E, Newton RR, Romney S, et al. Measuring the severity of child maltreatment. *Child Abuse and Neglect*. 2005; 29(5):553–73. <https://doi.org/10.1016/j.chiabu.2003.08.010> PMID: 15970325
22. Gillingham P. Predictive risk modelling to prevent child maltreatment: insights and implications from Aotearoa/New Zealand. *Journal of Public Child Welfare*. 2017; 11(2):150–65.
23. Keddell E. The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy*. 2015; 35(1):69–88.
24. Fleming M, Fitton CA, Steiner MFC, McLay JS, Clark D, King A, et al. Educational and health outcomes of children and adolescents receiving antiepileptic medication: Scotland-wide record linkage study of 766 244 schoolchildren. *BMC Public Health*. 2019; 19(1):595. <https://doi.org/10.1186/s12889-019-6888-9> PMID: 31101093
25. Fleming M, Fitton CA, Steiner MF, McLay JS, Clark D, King A, et al. Educational and health outcomes of children treated for asthma: Scotland-wide record linkage study of 683 716 children. *European Respiratory Journal*. 2019;54(3).

26. Gillies MB, Bowen JR, Patterson JA, Roberts CL, Torvaldsen S. Educational outcomes for children with cerebral palsy: a linked data cohort study. *Developmental Medicine and Child Neurology*. 2018; 60(4):397–401. <https://doi.org/10.1111/dmcn.13651> PMID: 29278268
27. Bell MF, Bayliss DM, Glauert R, Harrison A, Ohan JL. Chronic illness and developmental vulnerability at school entry. *Journal of Paediatrics*. 2016;137(5). <https://doi.org/10.1542/peds.2015-2475> PMID: 27244787
28. Kull MA, Coley RL. Early physical health conditions and school readiness skills in a prospective birth cohort of US children. *Social Science & Medicine*. 2015; 142:145–53.
29. Taras H, Potts-Datema W. Chronic health conditions and student performance at school. *Journal of School Health*. 2005; 75(7):255–66. <https://doi.org/10.1111/j.1746-1561.2005.00034.x> PMID: 16102088
30. Sturdy P, Bremner S, Harper G, Mayhew L, Eldridge S, Eversley J, et al. Impact of asthma on educational attainment in a socioeconomically deprived population: a study linking health, education and social care datasets. *PLoS One*. 2012; 7(11):e43977. <https://doi.org/10.1371/journal.pone.0043977> PMID: 23155367
31. Crump C, Rivera D, London R, Landau M, Erlendson B, Rodriguez E. Chronic health conditions and school performance among children and youth. *Annals of epidemiology*. 2013; 23(4):179–84. <https://doi.org/10.1016/j.annepidem.2013.01.001> PMID: 23415278
32. Green MJ, Kariuki M, Dean K, Laurens KR, Tzoumakis S, Harris F, et al. Childhood developmental vulnerabilities associated with early life exposure to infectious and noninfectious diseases and maternal mental illness. *Journal of Child Psychology and Psychiatry*. 2018; 59(7):801–10. <https://doi.org/10.1111/jcpp.12856> PMID: 29278269
33. Köhler-Forsberg O, Sørensen HJ, Nordentoft M, McGrath JJ, Benros ME, Petersen L. Childhood infections and subsequent school achievement among 598,553 Danish children. *The pediatric infectious disease journal*. 2018; 37(8):731–7. <https://doi.org/10.1097/INF.0000000000001869> PMID: 29278614
34. Azzam N, Oei J-L, Adams S, Bajuk B, Hilder L, Mohamed A-L, et al. Influence of early childhood burns on school performance: an Australian population study. *Archives of disease in childhood*. 2018; 103(5):444–51. <https://doi.org/10.1136/archdischild-2017-313355> PMID: 29187346
35. Sesko AM, Choe JC, Vitale MA, Ugwonali O, Hyman JE. Pediatric orthopaedic injuries: the effect of treatment on school attendance. *Journal of pediatric orthopaedics*. 2005; 25(5):661–5. <https://doi.org/10.1097/01.bpo.0000164869.35179.0f> PMID: 16199951
36. Gabbe BJ, Brooks C, Demmler JC, Macey S, Hyatt MA, Lyons RA. The association between hospitalisation for childhood head injury and academic performance: evidence from a population e-cohort study. *J Epidemiol Community Health*. 2014; 68(5):466–70. <https://doi.org/10.1136/jech-2013-203427> PMID: 24419234
37. Butler DC, Thurecht L, Brown L, Konings P. Social exclusion, deprivation and child health: a spatial analysis of ambulatory care sensitive conditions in children aged 0–4 years in Victoria, Australia. *Social Science & Medicine*. 2013; 94:9–16. <https://doi.org/10.1016/j.socscimed.2013.06.029> PMID: 23931940
38. Procter AM, Pilkington RM, Lynch JW, Smithers LG, Chittleborough CR. Potentially preventable hospitalisations in children: a comparison of definitions. *Archives of Disease in Childhood*. 2020; 105(4):375–81. <https://doi.org/10.1136/archdischild-2019-316945> PMID: 31666242
39. Scanlon MC, Harris JM, Levy F, Sedman A. Evaluation of the agency for healthcare research and quality pediatric quality indicators. *Pediatrics*. 2008; 121(6):e1723–e31. <https://doi.org/10.1542/peds.2007-3247> PMID: 18474532
40. Gill PJ. Developing paediatric quality indicators for UK general practice: Oxford University, UK; 2013.
41. Anderson P, Craig E, Jackson G, Jackson C. Developing a tool to monitor potentially avoidable and ambulatory care sensitive hospitalisations in New Zealand children. *NZ Med J*. 2012; 125(1366):25–37. PMID: 23254524
42. Falster K, Banks E, Lujic S, Falster M, Lynch J, Zwi K, et al. Inequalities in pediatric avoidable hospitalizations between Aboriginal and non-Aboriginal children in Australia: a population data linkage study. *BMC pediatrics*. 2016; 16(1):1–12.
43. National Assessment Program Literacy And Numeracy (NAPLAN) 2021 [Available from: <https://www.nap.edu.au/home>].
44. Australia Co. Australian Early Development Census 2021 [Available from: <https://www.aedc.gov.au/>].
45. Warren D, Haisken-DeNew JP. Early bird catches the worm: The causal impact of pre-school participation and teacher qualifications on Year 3 National NAPLAN Cognitive Tests. The University of Melbourne; 2013.

46. Miller PW, Voon D. Government Versus Non-Government Schools: A Nation-Wide Assessment Using Australian NAPLAN Data. *Australian Economic Papers*. 2012; 51(3):147–66.
47. Henry KL, Knight KE, Thornberry TP. School disengagement as a predictor of dropout, delinquency, and problem substance use during adolescence and early adulthood. *Journal of Youth and Adolescence*. 2012; 41(2):156–66. <https://doi.org/10.1007/s10964-011-9665-3> PMID: 21523389
48. Archambault I, Janosz M, Morizot J, Pagani L. Adolescent behavioral, affective, and cognitive engagement in school: Relationship to dropout. *Journal of School Health*. 2009; 79(9):408–15. <https://doi.org/10.1111/j.1746-1561.2009.00428.x> PMID: 19691715
49. Hancock KJ, Shepherd CC, Lawrence D, Zubrick SR. Student attendance and educational outcomes: Every day counts Canberra: Canberra: Department of Education, Employment and Workplace Relations; 2013 [Available from: <https://www.telethonkids.org.au/globalassets/media/documents/research-topics/student-attendance-and-educational-outcomes-2015.pdf>].
50. Ladwig JG, Luke A. Does improving school level attendance lead to improved school level achievement? An empirical study of indigenous educational policy in Australia. *The Australian Educational Researcher*. 2014; 41(2):171–94.
51. Gnanamanickam ES, Nguyen H, Armfield JM, Doidge JC, Brown DS, Preen DB, et al. Hospitalizations among children involved in the child protection system: A long-term birth cohort study from infancy to adulthood using administrative data. *Child Abuse and Neglect*. 2020; 107:104518. <https://doi.org/10.1016/j.chiabu.2020.104518> PMID: 32652507
52. Neil AL, Islam F, Kariuki M, Laurens KR, Katz I, Harris F, et al. Costs for physical and mental health hospitalizations in the first 13 years of life among children engaged with Child Protection Services. *Child Abuse and Neglect*. 2020; 99:104280. <https://doi.org/10.1016/j.chiabu.2019.104280> PMID: 31783310
53. Oh DL, Jerman P, Silvério Marques S, Koita K, Purewal Boparai SK, Burke Harris N, et al. Systematic review of pediatric health outcomes associated with childhood adversity. *BMC pediatrics*. 2018; 18(1):83. <https://doi.org/10.1186/s12887-018-1037-7> PMID: 29475430
54. Guthridge SL, Ryan P, Condon JR, Moss JR, Lynch JMMjoA. Trends in hospital admissions for conditions associated with child maltreatment, Northern Territory, 1999–2010. *Medical Journal of Australia*. 2014; 201(3):162–6. <https://doi.org/10.5694/mja14.00015> PMID: 25128952
55. Bell J, Lingam R, Wakefield CE, Fardell JE, Zeltzer J, Hu N, et al. Prevalence, hospital admissions and costs of child chronic conditions: A population-based study. *Journal of Paediatrics and Child Health*. 2020; 56(9):1365–70. <https://doi.org/10.1111/jpc.14932> PMID: 32502332
56. Cohen E, Berry JG, Camacho X, Anderson G, Wodchis W, Guttman A. Patterns and costs of health care use of children with medical complexity. *Paediatrics*. 2012; 130(6):e1463–e70. <https://doi.org/10.1542/peds.2012-0175> PMID: 23184117
57. Fang X, Fry DA, Brown DS, Mercy JA, Dunne MP, Butchart AR, et al. The burden of child maltreatment in the East Asia and Pacific region. *Child Abuse and Neglect*. 2015; 42:146–62. <https://doi.org/10.1016/j.chiabu.2015.02.012> PMID: 25757367
58. Leckning B, He VY, Condon JR, Hirvonen T, Milroy H, Guthridge S. Patterns of child protection service involvement by Aboriginal children associated with a higher risk of self-harm in adolescence: a retrospective population cohort study using linked administrative data. *Child Abuse and Neglect* 2021; 113:104931. <https://doi.org/10.1016/j.chiabu.2021.104931> PMID: 33461112
59. Warmingham JM, Handley ED, Rogosch FA, Manly JT, Cicchetti D. Identifying maltreatment subgroups with patterns of maltreatment subtype and chronicity: A latent class analysis approach. *Child Abuse and Neglect*. 2019; 87:28–39. <https://doi.org/10.1016/j.chiabu.2018.08.013> PMID: 30224068
60. Rivera PM, Fincham FD, Bray BC. Latent classes of maltreatment: A systematic review and critique. *Child Maltreatment*. 2018; 23(1):3–24. <https://doi.org/10.1177/1077559517728125> PMID: 28875728
61. McGuire A, Cho B, Huffhines L, Gusler S, Brown S, Jackson Y. The relation between dimensions of maltreatment, placement instability, and mental health among youth in foster care. *Child Abuse and Neglect*. 2018; 86:10–21. <https://doi.org/10.1016/j.chiabu.2018.08.012> PMID: 30248493
62. McGuire A, Jackson Y. Dimensions of maltreatment and academic outcomes for youth in foster care. *Child Abuse and Neglect*. 2018; 84:82–94. <https://doi.org/10.1016/j.chiabu.2018.07.029> PMID: 30071396
63. Guthridge SL, Ryan P, Condon JR, Bromfield LM, Moss JR, Lynch JW. Trends in reports of child maltreatment in the Northern Territory, 1999–2010. *Medical Journal of Australia*. 2012; 197(11–12):637–41. <https://doi.org/10.5694/mja12.10133> PMID: 23230935
64. Population: Northern Territory Government; 2021 [Available from: <https://nteconomy.nt.gov.au/population>].
65. Health Alo, Welfare. Profile of Indigenous Australians. Canberra: AIHW; 2021.

66. ABS. Australian Demographic Statistics, Jun 2016. Canberra: ABS. 2016(ABS cat. No. 3101.0).
67. Gracey M, King M. Indigenous health part 1: determinants and disease patterns. *The Lancet*. 2009; 374(9683):65–75. [https://doi.org/10.1016/S0140-6736\(09\)60914-4](https://doi.org/10.1016/S0140-6736(09)60914-4) PMID: 19577695
68. Bailie RS, Wayte KJ. Housing and health in Indigenous communities: Key issues for housing and health improvement in remote Aboriginal and Torres Strait Islander communities. *Australian Journal of Rural Health*. 2006; 14(5):178–83. <https://doi.org/10.1111/j.1440-1584.2006.00804.x> PMID: 17032292
69. Zhao Y, You J, Wright J, Guthridge SL, Lee AH. Health inequity in the northern territory, Australia. *International Journal for Equity in Health*. 2013; 12(1):1–8. <https://doi.org/10.1186/1475-9276-12-79> PMID: 24034417
70. O'Donnell M, Nassar N, Leonard H, Jacoby P, Mathews R, Patterson Y, et al. Rates and types of hospitalisations for children who have subsequent contact with the child protection system: a population based case-control study. *Journal of Epidemiology and Community Health*. 2010; 64(9):784–8. <https://doi.org/10.1136/jech.2009.093393> PMID: 19778908
71. Child protection Australia: 2017–18 [press release]. Australian Government 2019.
72. Möller H, Falster K, Ivers R, Falster M, Randall D, Clapham K, et al. Inequalities in hospitalized unintentional injury between Aboriginal and non-Aboriginal children in New South Wales, Australia. *American Journal of Public Health*. 2016; 106(5):899–905. <https://doi.org/10.2105/AJPH.2015.303022> PMID: 26890169
73. Guenther J, Lowe K, Burgess C, Vass G, Moodie N. Factors contributing to educational outcomes for First Nations students from remote communities: A systematic review. *The Australian Educational Researcher*. 2019; 46(2):319–40.
74. Education NGDo. Enrolment and attendance 2020 [Available from: <https://education.nt.gov.au/statistics-research-and-strategies/enrolment-and-attendance>].
75. Australian Curriculum AaRA. Measurement Framework for Schooling in Australia 2020 [Available from: <https://www.acara.edu.au/reporting/measurement-framework-for-schooling-in-australia>].
76. Australia Co. National Agreement on Closing The Gap 2021 [Available from: <https://www.closingthegap.gov.au/>].
77. Australia's children Canberra: AIHW; 2020 [Available from: <https://www.aihw.gov.au/reports/children-youth/australias-children>].
78. Inquest into the deaths of Fionica Yarranganlagi James, Keturah Cheralyn Mamarika and Layla Leering, (2020).
79. Government NT. Multi-Agency Community and Child Safety Framework 2020 [Available from: <https://tfhc.nt.gov.au/children-and-families/multi-agency-community-and-child-safety-framework#:~:text=The%20MACCSF%20was%20introduced%20in,communities%20with%20a%20twofold%20purpose%3A&text=improving%20the%20efficient%20and%20effective%20use%20of%20resources%20for%20communities>].
80. Stanley N, Humphreys C. Multi-agency risk assessment and management for children and families experiencing domestic violence. *Children and Youth Services Review*. 2014; 47:78–85.
81. He V, Guthridge S, Leckning B. From Birth to Five: A multiagency data-linkage study to inform a public health response to child protection in the Northern Territory. Darwin: Menzies School of Health Research. 2019.
82. DataLink S-N. SA-NT DataLink 2021 [Available from: <https://www.santdatalink.org.au/>].
83. Silburn S. Early pathways to school learning: Lessons from the NT data linkage study: Centre for Child Development and Education, Menzies School of Health Research; 2018.
84. ABS. Schools, 2020. Australian Bureau Statistics; 2020.
85. Anderson P, Craig E, Jackson G, Jackson C. Developing a tool to monitor potentially avoidable and ambulatory care sensitive hospitalisations in New Zealand children. *NZ Medical Journal*. 2012; 125(1366):25–37. PMID: 23254524
86. Government NT. Report Child Abuse 2021 [Available from: <https://nt.gov.au/law/crime/report-child-abuse>].
87. Kerr J. Policy Determination 5.2: Mandatory Reporting. 2018.
88. Hackeling G. Mastering Machine Learning with scikit-learn: Packt Publishing Ltd; 2017.
89. Grant RW, McCloskey J, Hatfield M, Uratsu C, Ralston JD, Bayliss E, et al. Use of Latent Class Analysis and k-Means Clustering to Identify Complex Patient Profiles. *JAMA Network Open*. 2020; 3(12): e2029068–e. <https://doi.org/10.1001/jamanetworkopen.2020.29068> PMID: 33306116
90. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 2011; 12:2825–30.

91. Alonso MC, Malpica JA, de Agirre AM, editors. Consequences of the Hughes phenomenon on some classification techniques. *Proceedings of the ASPRS 2001 Annual Conference*; 2011.
92. Bouveyron C, Girard S, Schmid C. High-dimensional data clustering. *Computational Statistics and Data Analysis*. 2007; 52(1):502–19.
93. Raftery AE, Dean N. Variable selection for model-based clustering. *Journal of the American Statistical Association*. 2006; 101(473):168–78.
94. Dy JG, Brodley CE. Feature selection for unsupervised learning. *Journal of Machine Learning Research*. 2004; 5(Aug):845–89.
95. Adams S, Beling PA. A survey of feature selection methods for Gaussian mixture models and hidden Markov models. *Artificial Intelligence Review*. 2019; 52(3):1739–79.
96. Law MH, Jain AK, Figueiredo MA, editors. Feature selection in mixture-based clustering. *NIPS*; 2002.
97. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018; 300:70–9.
98. Ringnér M. What is principal component analysis? *Nature Biotechnology*. 2008; 26(3):303–4. <https://doi.org/10.1038/nbt0308-303> PMID: 18327243
99. Kocev D, Slavkov I, Dzeroski S, editors. Feature ranking for multi-label classification using predictive clustering trees. *International Workshop on Solving Complex Machine Learning Problems with Ensemble Methods, in conjunction with ECML/PKDD*; 2013.
100. Fred AL, Jain AK. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2005; 27(6):835–50. <https://doi.org/10.1109/TPAMI.2005.113> PMID: 15943417
101. Lourenço A, Rota Bulò S, Rebagliati N, Fred AL, Figueiredo MA, Pelillo M. Probabilistic consensus clustering using evidence accumulation. *Machine Learning*. 2015; 98(1):331–57.
102. Team RC. R: a language and environment for statistical computing, version 3.5. 1 Vienna, Austria 2018 [Available from: <https://www.R-project.org/>].
103. Van Rossum G, Drake FL Jr. *Python reference manual*: Centrum voor Wiskunde en Informatica Amsterdam; 1995.
104. Gulli A, Pal S. *Deep learning with Keras*: Packt Publishing Ltd; 2017.
105. Leadership AlfTaS. *Spotlight: Attendance Matters*. 2019.
106. Cohen SD. *Applying the Science of Child Development in Child Welfare System 2016* [Available from: [www.developingchild.harvard.edu](http://www.developingchild.harvard.edu)].
107. Bullinger LR, Feely M, Raissian KM, Schneider W. Heed neglect, disrupt child maltreatment: A call to action for researchers. *International Journal on Child Maltreatment: Research, Policy and Practice*. 2020; 3(1):93–104.
108. Sonnenschein-Van Der Voort AM, Arends LR, de Jongste JC, Annesi-Maesano I, Arshad SH, Barros H, et al. Preterm birth, infant weight gain, and childhood asthma risk: a meta-analysis of 147,000 European children. *Journal of Allergy and Clinical Immunology*. 2014; 133(5):1317–29. <https://doi.org/10.1016/j.jaci.2013.12.1082> PMID: 24529685
109. Baraldi E, Filippone M. Chronic lung disease after premature birth. *New England Journal of Medicine*. 2007; 357(19):1946–55. <https://doi.org/10.1056/NEJMra067279> PMID: 17989387
110. Hafekost K, Lawrence D, O’Leary C, Bower C, O’Donnell M, Semmens J, et al. Maternal alcohol use disorder and subsequent child protection contact: A record-linkage population cohort study. *Child Abuse and Neglect*. 2017; 72:206–14. <https://doi.org/10.1016/j.chiabu.2017.08.010> PMID: 28823788
111. Leek L, Seneque D, Ward K. Parental drug and alcohol use as a contributing factor in applications to the Children’s Court for protection orders. *Children Australia*. 2009; 34(2):11–6.
112. Cooke L, Piers-Blundell A. Embedding Evidence-Based Practice into a Remote Indigenous Early Learning and Parenting Program: A Systematic Approach. *Literacy Education and Indigenous Australians*: Springer; 2019. p. 185–201.
113. Kruske S, Belton S, Wardaguga M, Narjic C. Growing up our way: the first year of life in remote Aboriginal Australia. *Qualitative Health Research*. 2012; 22(6):777–87. <https://doi.org/10.1177/1049732311432717> PMID: 22218266
114. Zhao Y, Wright J, Guthridge S, Lawton P. The relationship between number of primary health care visits and hospitalisations: evidence from linked clinic and hospital data for remote Indigenous Australians. *BMC Health Services Research*. 2013; 13(1):1–9. <https://doi.org/10.1186/1472-6963-13-466> PMID: 24195746
115. Segal L, Nguyen H, Gent D, Hampton C, Boffa J. Child protection outcomes of the Australian Nurse Family Partnership Program for Aboriginal infants and their mothers in Central Australia. *PLOS One*. 2018; 13(12):e0208764. <https://doi.org/10.1371/journal.pone.0208764> PMID: 30532276

116. Nguyen H, Zarnowiecki D, Segal L, Gent D, Silver B, Boffa J. Feasibility of implementing infant home visiting in a Central Australian Aboriginal community. *Prevention Science*. 2018; 19(7):966–76. <https://doi.org/10.1007/s11121-018-0930-5> PMID: 30054778
117. Kerr J. A descriptive analysis of the characteristics, seriousness and frequency of Aboriginal intimate partner violence in the Northern Territory, Australia: a strategy for targeting high harm cases. Unpublished Masters Thesis Oxford, UK: Oxford University 2016.
118. Lim KHA, McDermott K, Read DJ. Interpersonal violence and violent re-injury in the Northern Territory. *Australian Journal of Rural Health*. 2020; 28(1):67–73. <https://doi.org/10.1111/ajr.12590> PMID: 31970833
119. McCallum K, Waller L. Un-braiding deficit discourse in Indigenous education news 2008–2018: performance, attendance and mobility. *Critical Discourse Studies*. 2020:1–20.
120. Millier J. Moving away from the deficit perspective: Aboriginal child development in partnership with families. *Rural Society*. 2021; 30(1):1–14.
121. Fogarty W, Lovell M, Dodson M. Indigenous education in Australia: Place, pedagogy and epistemic assumptions. *UNESCO Observatory Refereed e-Journal*. 2015; 4:1–21.
122. Wilson B. A share in the future: Review of Indigenous education in the Northern Territory: Education Business; 2014.
123. Benveniste T, Guenther J, Dawson D, Rainbird S. Out of Sight, Out of Mind? Bringing Indigenous Parent-Boarding School Communication to Light. Australian Association for Research in Education. 2014.
124. Fogarty W, Lovell M, Dodson M. A view beyond review: Challenging assumptions in Indigenous education development. *UNESCO Observatory Multi-Disciplinary Journal in the Arts*. 2015.
125. Altman J, Kerins S, Fogarty W. Why the Northern Territory Government needs to support Outstations/Homelands in the Aboriginal, Northern Territory and National Interest-The Importance of Supporting Outstations. 2018.
126. NT Education Engagement Strategy 2022–2031: NT Government; 2022 [Available from: <https://education.nt.gov.au/statistics-research-and-strategies/education-engagement-strategy>].
127. Rioux J, Smith G. Both-Ways science education: Place and context. *Learning Communities*. 2019.
128. Van Gelderen B. 'Growing our own': A 'two way', place-based approach to Indigenous initial teacher education in remote Northern Territory. *Australian and International Journal of Rural Education*. 2017; 27(1):14–28.
129. Neesham G, Garnham AP. Success story: Clontarf Foundation promotes education, life-skills and employment prospects through Australian Rules Football. *British Journal of Sports Medicine*. 2012; 46(13):898–9. <https://doi.org/10.1136/bjsports-2012-091686> PMID: 23007176
130. Rawlinson C. Stronger Smarter Sisters take on education challenge. ABC Local 2011.
131. Guenther J, editor *Conspiring to inspire community aspirations through remote education*. Australian Association for Research in Education; 2015; Fremantle: Australian Association for Research in Education Annual Conference, Fremantle.
132. Remote School Attendance Program: Papulu Appar-Kari Aboriginal Corporation; 2022 [Available from: <https://papak.com.au/services/yellow-shirt-crew>].
133. Program Results in Record Attendance at Wadeye School: Batchelor Institute; 2016 [Available from: <https://www.batchelor.edu.au/portfolio/program-results-in-record-attendance-at-wadeye-school/>].
134. He VY, Leckning B, Malvaso C, Williams T, Liddle L, Guthridge S. Opportunities for prevention: a data-linkage study to inform a public health response to youth offending in the Northern Territory, Australia. *BMC Public Health*. 2021; 21(1):1–14.
135. developers sl. K-means 2021 [Available from: <https://scikit-learn.org/stable/modules/clustering.html#k-means>].
136. Peterson A, Joseph J, M F. New directions in child abuse and neglect research. 2014. Report No.: 0309285151.
137. Child NSCoTD. The science of neglect: The persistent absence of responsive care disrupts the developing brain: working Paper No. 12.: Harvard University; 2012 [Available from: [www.developingchild.harvard.edu](http://www.developingchild.harvard.edu)].
138. Shonkoff JP, Levitt P, Boyce T, Cameron J, Duncan G, Fox N, et al. Persistent fear and anxiety can affect young children's learning and development: Working Paper No. 9. National Scientific Council on the Developing Child. 2010:1–13.
139. Abdi H, Williams L. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2010; 2(4):433–59.
140. Song M, Yang H, Siadat SH, Pechenizkiy M. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. *Expert Systems with Applications*. 2013; 40(9):3722–37.