

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Can routinely collected primary healthcare data be used to assess Aboriginal children's health and wellbeing longitudinally? A retrospective analysis of electronic medical records from an Aboriginal community-controlled health service in Central Australia

Catherine Lloyd-Johnsen^{1,2*}, John Boffa³, Vahab Baghbanian^{3,4}, Rachel Walpole³, Shuaijun Guo¹, Sandra Eades^{2,5}, Anita D'Aprano^{1,2}, and Sharon Goldfeld^{1,2}

Submission History

Submitted:	05/12/2024
Accepted:	01/05/2025
Published:	01/07/2025

¹Centre for Community Child Health, Royal Children's Hospital, Murdoch Children's Research Institute, Melbourne, VIC, Australia

²Department of Paediatrics, University of Melbourne, Melbourne, VIC, Australia

³Central Australian Aboriginal Congress, Alice Springs, NT, Australia

⁴Adelaide Health Technology Assessment, University of Adelaide, Adelaide SA, Australia

⁵Melbourne School of Population and Global Health, University of Melbourne, Melbourne, VIC, Australia

Abstract

Introduction

Electronic medical records (EMR) are an essential tool in modern healthcare, providing a centralised source of patient information. Longitudinal analysis of EMRs can identify opportunities for targeted interventions to improve health outcomes for children. However, the research value of EMRs is contingent on data quality and completeness.

Methods

This retrospective cohort study used deidentified EMRs from all Aboriginal children born in 2015 who attended an Aboriginal-controlled health service in Central Australia over a 5-year period. The purpose of this study was to demonstrate the utility of EMRs in longitudinal research via presentation of three case study example analyses, and to evaluate the quality of the extracted dataset.

Results

EMRs of 319 Aboriginal children (48.9% girls, 51.1% boys) were included in the analysis. These children visited the service an average of 19.9 times (min 2 - max 102). Attendance rates for routine well-child check-ups were highest at 0 to 8 weeks and 4 years of age (37.3% and 40.1% respectively). Among 12-month-olds with recorded haemoglobin levels, 43% were anaemic. Weight-for-age medians were comparable to World Health Organization (WHO) growth standards until 12 months age, thereafter Aboriginal girls tended to weigh more overtime. Data completeness varied: key variables (date of birth, sex and Aboriginal status) were 100% complete, while others like anthropometrics (up to 62.1%), birth weight (54.2%), gestational age (50.2%), and haemoglobin results (up to 34.1%) were less complete. Average accuracy (99.2%) and consistency of available data (100%) were high. However, crucial data on risk factors, maternal health, and family functioning were either not collected by the service, not provided to the service from external sources, or stored in inaccessible free-text fields.

Conclusions

Missing data were the greatest limiting factor for reporting on the health and development of these children. To reap the benefit of utilising EMRs for longitudinal research, the service should continue encouraging families to attend their child's routine health assessments in the first years of life. Setting key data variables as mandatory at each visit may also help increase data completeness over time.

Keywords

Aboriginal and Torres Strait Islander children; electronic health records; retrospective cohort; longitudinal; administrative data

*Corresponding Author:

Email Address: cat.lloydjohnsen@mcri.edu.au (Catherine Lloyd-Johnsen)

Introduction

Healthy development during the early years is critical as it forms the foundation for future health and wellbeing [1, 2]. Many Aboriginal and Torres Strait Islander (hereafter Aboriginal¹) children grow up in safe and nurturing environments, connected to culture and supported by caring relationships with their parents, extended family, and communities. These supportive relationships play a crucial role in their development and overall wellbeing. However, not all Aboriginal children have such a strong start to life. The cumulative effect of multiple adverse childhood experiences (ACEs) including poverty, maltreatment, family violence and dysfunction can negatively impact an individual's life-long health and wellbeing [3, 4]. Health disparities between Aboriginal and non-Aboriginal children are well documented. Aboriginal children have higher rates of chronic diseases, are more likely to live in sub-optimal housing conditions, face challenges in accessing quality education and are disproportionately affected by poverty [5]. While substantial investment is being made to understand the origins of these disparities, the gap continues to widen [6, 7].

Policy planning and decision-making processes aimed at addressing this gap are best supported by robust evidence [8, 9]. However, Aboriginal researchers have critiqued the presentation of over-simplified, heavily processed national or state level aggregate data as deficit focused and decontextualised [10]. These official statistics mask geographic differences and often do not reflect the diverse realities of communities. Therefore, access to high-quality disaggregated regional data is necessary to better understand the specific needs and aspirations of local Aboriginal communities, and to develop and evaluate place-based solutions that promote their health, wellbeing and self-determination [7, 11].

The Central Australian Aboriginal Congress Aboriginal Corporation (hereafter Congress) has been working in partnership with local communities to deliver primary health care to Aboriginal people since 1973. A large Aboriginal-controlled health service based in Alice Springs, Congress services the Central Australian Aboriginal population through 16 clinics, social and preventive services and programs and advocacy on the social determinants of health. The service has a long-term goal to support the development of longitudinal research focusing on early childhood [12]. Partnering with the Murdoch Children's Research Institute and the University of Melbourne, Congress set out to investigate the feasibility and acceptability of establishing a prospective cohort study of young Aboriginal children living in and around Alice Springs. A cohort study is a type of observational study where the same group of individuals are followed over a period of time [13]. By collecting data at multiple time points, researchers can identify trends, detect patterns, and assess the impact of different variables on the observed outcomes [14].

Electronic capture of clinical information across the continuum of care is now widely used in primary care. An

electronic medical record (EMR) is a digital version of a client's file that includes their medical history, diagnoses, medications, treatment plans and other clinical notes as recorded by clinicians and staff during healthcare encounters. Service providers can use these records to make informed decisions about health care delivery. EMRs have been used by Congress since 2007 to improve health care quality, health service planning, and monitor system performance. The secondary use of these data for research purposes is rapidly expanding [15, 16]. Integrating prospectively collected cohort data with existing EMRs has the potential to maximise value without imposing additional burden on cohort participants [17, 18]. Through the longitudinal analysis of existing EMRs, Congress can demonstrate the effectiveness of its programs over time and increase its understanding of what outcomes are being achieved for its young clients. However, the suitability of using Congress EMRs for longitudinal research will greatly depend on the quality and completeness of the records [19]. For these reasons, we aimed to (1) demonstrate the potential utility of Congress-held EMRs for longitudinal research and (2) assess the quality and completeness of an extracted deidentified dataset.

Methods

Study design

This retrospective cohort study was designed as a partnership project between Congress, the Murdoch Children's Research Institute (MCRI) and the University of Melbourne. It is part of the Atyepe-atyeppe Iwerre Ampe-ke (Healthy Journey for Kids: Feasibility Study).

Study population

Alice Springs Hospital is the primary public healthcare facility for the Central Australia region and serves as a major birthing centre. In 2015, the hospital recorded 668 births, 47% of which were Aboriginal babies [20]. Given the extensive outreach of Congress, it is likely that a significant proportion of these families routinely utilise their services. This retrospective cohort included all Aboriginal children born in 2015 who attended the service more than once over the study period (January 1, 2015, to December 31, 2020) and who were residents of Central Australia (n = 319).

Data extraction

A decision was made to focus on high value variables or proxy measures that are known to be important in the longitudinal analysis of Aboriginal child health and wellbeing, as detailed in a published review of the international literature [21]. The complete list of variables requested from Congress is presented in Appendix 1. Not all of the requested items were available. Data were extracted from the clinical information system, Communicare (Telstra Health Communicare Systems, Perth, WA, Australia), in May 2022 using an automated script developed by the Congress Data Manager.

¹We respectfully use the term 'Aboriginal' to refer to the diverse Aboriginal, Aboriginal and Torres Strait Islander, and Torres Strait Islander peoples whose unceded lands and waters constitute present day Australia. We acknowledge that as a stand-alone term it is not inclusive of Torres Strait Islander peoples. However, this choice of terminology reflects the Central Australian preference to identify in this way.

All Aboriginal children born in 2015 with a minimum of two recorded visits (or episodes of care²) were eligible for inclusion in the study. To ensure confidentiality, the Congress Data Manager removed client identifiers, retaining only sex and month/year of birth. A new unique study identifier was created and assigned to each child. The deidentified dataset was securely sent to the external research team via an encrypted file. The data were then imported into STATA18 (Statacorp, College Station, Texas, USA) for cleaning and analysis. Appendix 2 details the extracted variables by data collection time point. Appendix 3 details the steps taken to prepare the dataset for analysis.

Data quality assessment

A data quality assessment was performed. Based on the published literature, we adapted a framework used by Liaw et al. (2011) to assess the data [22]. This included focus on completeness, correctness, and consistency, all expressed as percentages.

Completeness was defined as the availability of at least one record per child for each variable. The denominator was the total number of children born in 2015 with a minimum of two episodes of care ($n = 319$). We calculated the percentage of non-missing values over time to inform suitability for longitudinal analysis.

Correctness was defined as a valid and appropriate record, free from qualitative or quantitative error. To assess correctness, we looked for values that were either: (a) not biologically plausible, (b) not following a logical chronological sequence, (c) inconsistent in their measured units, (d) inconsistent across duplicate records for an individual client or (e) outliers (± 3 SDs). The correctness rate was defined as the number of records free from error divided by the total number of available records. This assessment relied solely on internal validation within the dataset, without external medical record review or manual validation.

Consistency focused on how closely the dataset aligns over time or between departments within the service, such as the use of uniform data types and formats (e.g., integer, string, date) with consistent labels or coding (e.g., gestation vs gestational age). We kept detailed field notes to document workflow data cleaning decisions during the data quality assessment. Data triangulation between multiple sources (e.g., comparing EMR to paper-based medical files) was outside the scope of approval granted for this study.

Data analysis

To illustrate the potential application of Congress EMRs in longitudinal research, the extracted data were analysed using descriptive statistics and presented across three example case studies on topics of particular interest to the service. Our focus was to highlight the feasibility of using the extracted data for longitudinal analysis, rather than outcomes reported.

²In this study we defined an "episode of care" as contact between an individual client and the health service, with one or more staff, to provide health care (e.g. for sickness, injury, counselling, health education, screening) across a single day.

The Northern Territory (NT) 'Healthy Under 5 Kids' (HU5K) program was designed to monitor the health of remote Aboriginal children at key age milestones between birth and 5 years of age [23]. It follows a schedule of well-child health checks at key developmental milestones, including assessments at around 10 days, 4 weeks, 8 weeks, 4 months, 6 months, 9 months, 12 months, 18 months, 2 years, 2 1/2 years, 3 years, 3 1/2 years, 4 years and 4 1/2 years of age. It also provides a standardised template for comprehensive well-child health assessments for remote NT staff to use. Congress elected to use the Medicare Benefits Schedule (MBS) Item 715, the Aboriginal and Torres Strait Islander Peoples Health Assessment, for children presenting to their clinics rather than the HU5K templates, as there is considerable overlap between the two and use of the former facilitates access to the Medicare rebate. The first example analysis (Case Study 1) maps the cohort's recorded attendance at these routine check-ups over the study period with some modifications. Participation rates were based on the number of children with any data recorded at each check-up, expressed as a percentage of the number of children eligible to attend in the cohort.

The second example analysis (Case Study 2) presents physical growth from birth to 4 years of age. Congress staff measure a child's weight measured to the nearest 0.1 kg using standard digital scales and height/length³ measured to the nearest 0.1 cm using stadiometers at each routine check-up. In this analysis, we plotted weight (kg), height (cm) and calculated body mass index (BMI, kg/m²) profiles for boys and girls with complete data from birth to 4 years of age. Weight-for-age and height-for-age percentiles were calculated in STATA and plotted against matched World Health Organization (WHO) Child Growth Standards [24].

The final example analysis (Case Study 3) looks at the prevalence of anaemia over time. Haemoglobin (Hb) levels are routinely measured on a sample of capillary blood taken by finger prick or heel prick, every 6 months from 6 months of age. In this analysis, children aged six to 11 months with a Hb reading below 105g/L were considered to be anaemic. Children aged 1 to 4 years, with a Hb reading below 110g/L, and children aged 5 years with a Hb reading below 115g/L were also considered to be anaemic, as per current guidelines [25].

Demographics and other clinical characteristics of the cohort were also evaluated and described where possible. Aboriginality was predefined by Congress in their clinical information system. Free-text fields were not made available to the research team as anonymity could not easily be assured. Frequency and percentage were used to summarise categorical variables. The mean and standard deviation (SD) or median and interquartile range (IQR) were used to summarise continuous variables. STATA18 was used for all statistical analyses. A brief description of the cohort precedes the case study examples.

Results

The full cohort consisted of 319 Aboriginal children born in 2015 (Girls: $n = 156$, 48.9%, Boys: $n = 163$, 51.1%) who were recorded as residents of Central Australia over the

³The term 'height' will be used throughout this paper to describe both height and length.

study period, including 315 singleton infants and two sets of twins. Hospital of birth was only available for 35.1% of the cohort ($n = 112$). The vast majority of children with data on hospital of birth were born locally at the Alice Springs Hospital ($n = 100$, 89.3%).

A valid birthweight was recorded for 173 children (54.2%) in the cohort (Table 1). The average birthweight was 3.2kg, which is only 100 grams lower than the national average of 3.3kg in 2015 [26]. Girls had lower average birth weights compared to boys: 3.1kg versus 3.3kg and appear to be categorised as low birth weight (≤ 2.5 kg) more frequently than boys ($n = 18$, 10.4% versus $n = 7$, 4%). Data on recumbent length at birth was only available for 146 (45.8%) children. Girls were twice as likely to be born preterm (< 37 weeks gestation) compared to boys (Girls, $n = 15$, 9.4% versus Boys, $n = 7$, 4.4%).

Breast feeding rates were high in our cohort with 85.3% of mothers (87 out of 102 with data) reported to have exclusively breastfed their newborn infants at hospital discharge. We found that exclusive breastfeeding rates dropped to 55.7% by 3 months of age ($n = 54$ out of 94 with follow-up data), similar to rates among Aboriginal infants reported elsewhere [27].

The service collects some socio-demographic data but not routinely. Most often this information is stored in free-text clinical notes. Thus, the research team did not have access

to important socio-demographic factors such as: maternal education, occupation, income, and/or place of residence. However, the categorical variable “main carer” was collected ad hoc over time in a small sub-sample of the cohort ($n = 69$, 21.6%). The majority of those children were cared for by their mother ($n = 60$, 86.9%). The rest were cared for by their grandparents or others ($n = 6$, 8.7%), both parents ($n = 2$, 2.9%) or by their father ($n = 1$, 1.5%).

Data on housing type was also collected at random intervals over the study period, limiting our ability to examine change over time. Of the 113 families (35.4%) with available data on housing type, the largest proportion ($n = 54$, 47.8%) reported living in public housing, followed by private houses or units ($n = 45$, 39.8%). Almost nine per cent ($n = 10$) had no fixed address and were temporarily staying with family or friends, while over three per cent ($n = 4$) were living in supported or crisis accommodation.

Data coverage of household size (number of bedrooms) and occupancy (number of adults and children) was inconsistent, limiting our ability to quantitatively describe the housing situation of families in our cohort over time. Using the first recorded observation (0-5 years of age), we found 81.6 per cent of families (80 out of 98) lived in homes with 3 bedrooms or less, and the median number of household occupants was six (three adults and three children). Congress staff asked

Table 1: Birth details of retrospective cohort ($n = 319$ born in 2015)

	N	%	MEAN	SD
Sex				
Missing	0	0		
Girls	156	48.9		
Boys	163	51.1		
Birth weight (kg)				
All children	173	54.2	3.2	0.674
Missing	146	45.8		
Girls	93		3.1	0.748
Boys	80		3.3	0.555
Low birth weight (Less than 2.5kg)	25	14.5	2.1	0.4
Normal birth weight (2.5-3.9kg)	129	74.6	3.3	0.4
Macrosomic (4kg and above)	19	10.9	4.3	0.3
Birth length (cm)				
All children	146	45.8	49.3	3.57
Missing	173	54.2		
Girls	82		48.4	3.89
Boys	64		50.4	2.77
Head circumference (cm)				
All children	143	44.8	34.4	2.47
Missing	176	55.2		
Girls	79		34.2	2.8
Boys	64		34.7	1.9
Gestational age at birth (weeks)				
All children	160	50.2	38.5	2.3
Missing	159	49.8		
Early/moderate preterm (≤ 33 weeks)	7	4.4	30.9	3.1
Late preterm (34–36 weeks)	15	9.4	35.6	0.9
Full term (37 weeks and above)	138	86.2	39.3	1.2

caregivers (67.7%, $n=216$) about potential child exposure to environmental tobacco smoke (ETS) at least once over the study period. The vast majority ($n=186$, 86.1%) answered "No". For the few families with serial data, ETS exposure was transient over time but less common in the first year of life.

Data on languages spoken was only available for a sub-sample of families in our cohort (97 responses to number of languages spoken and 106 responses to main language spoken at home). More than 64 per cent of the children ($n=68$) came from homes in which one or more Aboriginal languages were spoken. English was the main language spoken in 38 households. Following this, Arrernte was the next most frequently spoken language in 33 households (31.1%). Other main languages spoken included Luritja ($n=9$, 8.5%), Pitjanjatjara ($n=8$, 7.5%), Walpiri ($n=8$, 7.5%), Anmatyerr ($n=4$, 3.8%) and Alyawarr ($n=2$, 1.9%).

About 15% of our cohort ($n=48$) were screened (at least once) for developmental status using the Ages and Stages Questionnaires[®] – Talking about Raising Aboriginal Kids (ASQ-TRAK) developmental screening tool, adapted and validated for use in this population [28]. Data on participation in childcare, kindergarten, or other early childhood education programs was scant. Of those children with available data ($n=123$), 63.4% ($n=78$) were noted as having attended preschool or ECE program by 4 and 5 years of age. Forty-three of these children attended the Congress Child Health and Development (CHAD) which opened in 2017.

CASE STUDY 1: Attendance at routine check-ups

The median number of visits to the service per child was 14, ranging from between 2 to 102 visits each. Over the study period, boys appear to visit the service more frequently than girls (22 versus 17 visits on average). Overall, 79.6% ($n=254$) of the cohort attended at least one of their scheduled routine check-ups. Only twelve children from our cohort attended routine check-ups every year from 1 to 4 years of age. Over

a third of our cohort attended the 4-year routine check-up ($n=128$, 40.1%) (Figure 1) while the 5-year-old check-up had the lowest attendance rate with only 9 children with recorded data ($n=9$, 2.8%).

CASE STUDY 2: Physical growth

Height and weight were serially measured over the study period at routine check-ups. There were 18 children (9 girls and 9 boys) with complete longitudinal height and weight data at birth, 6, 12, 18 months, 2, 3 and 4 years of age. Height and weight collected at 5 years of age were excluded due to the high proportion of missing data. The mean (SD) and median (IQR) weight and height for this sub-group of children are presented in Appendix 4. In this small sample, boys were heavier than girls until 3 years of age when the trend reversed with girls weighing slightly more on average.

Descriptive weight-for-age and height-for-age growth profiles are presented alongside WHO Growth Standards in Figure 2. No statistical comparisons were conducted. Median weight-for-age among Congress girls appeared comparable to the WHO reference population in the first 6 months of life. After that, Congress girls tended to weigh more than their age matched WHO peers. Congress boy's median weights were broadly similar to WHO growth standards across most timepoints, with minor fluctuations at 12 and 24 months of age. Congress boys were taller than Congress girls at all measured time points, with the largest difference observed at 18 months of age. The Congress sample of boys and girls appeared shorter than their respective WHO peers from 3 years of age onwards.

BMI was categorised into underweight ($\text{BMI} \leq 14.42$ for boys and $\text{BMI} \leq 14.18$ for girls), normal weight ($\text{BMI} 14.43\text{--}17.54$ for boys and $14.19\text{--}17.27$ for girls), overweight or obese ($\text{BMI} \geq 17.55$ for boys and $\text{BMI} \geq 17.28$ for girls) [29–31]. For the 175 four-year-olds with complete data (88 girls and 87 boys), we found fewer than 1 in 8 children ($n=22$, 12.6%) were categorised as underweight. When stratified by sex, the

Figure 1: Recorded attendance at routine check-ups at Congress ($n=319$)

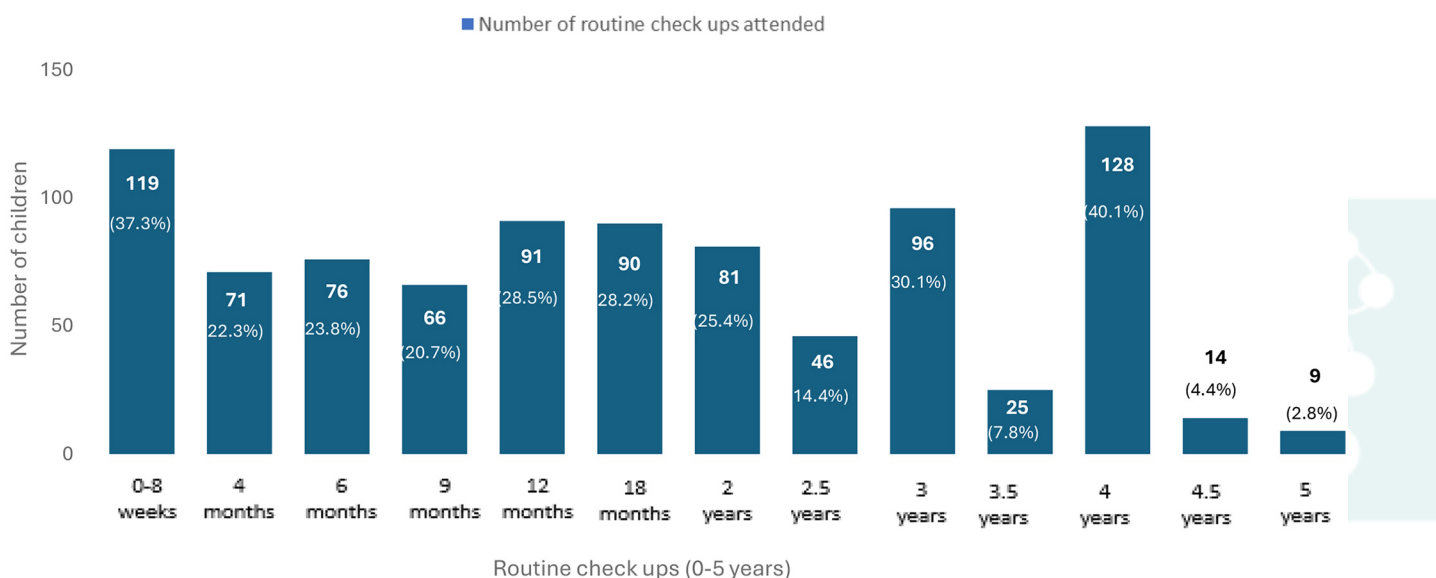
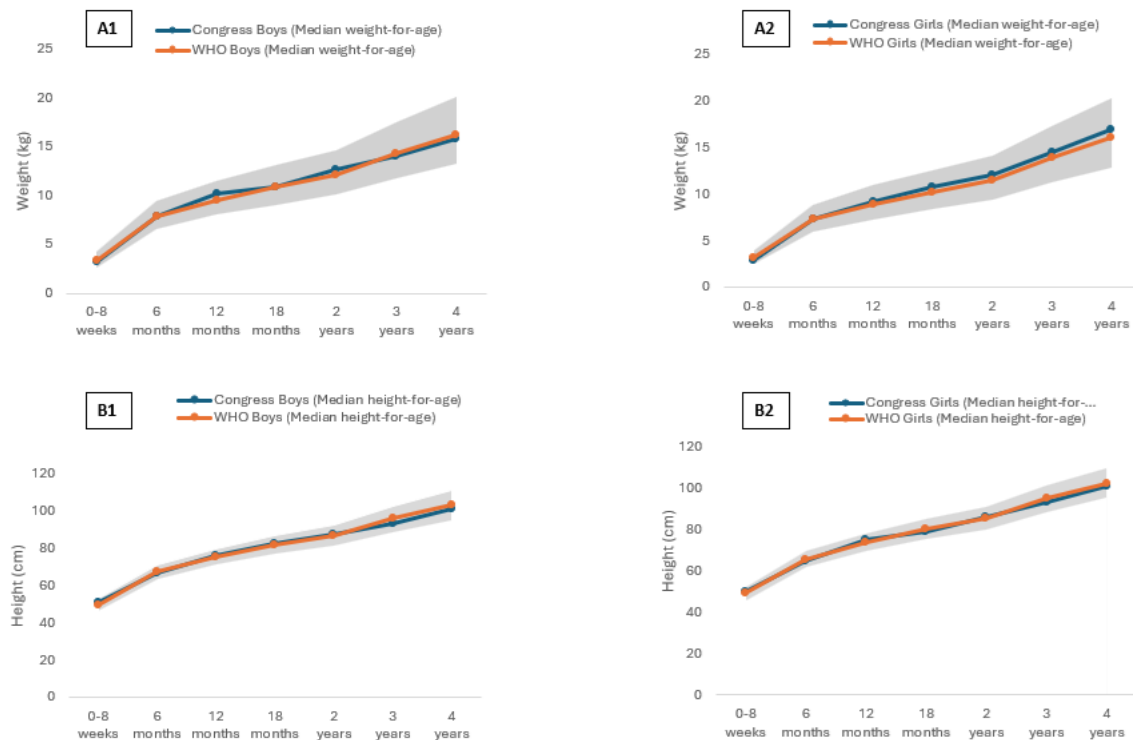


Figure 2: Growth trajectories for Congress children (n=18) over time compared to WHO Child Growth Standards



Growth trends (weight and height) for Congress boys and girls over time. (A1) Weight-for-age in Congress boys (n = 9). (A2) Weight-for-age in Congress girls (n = 9). (B1) Height-for-age in Congress boys (n = 9). (B2) Height-for-age in Congress girls (n = 9). All lines represent median values. The blue line is the 50th centile for Congress children over time, the orange dotted line is the 50th centile according to the WHO Child Growth Standards. The grey areas are the WHO Child Growth Standards 5th to 95th centiles for weight and height (by sex).

proportion of underweight girls is slightly higher than boys (n = 12, 13.6% vs n = 10, 11.49%). Most four-year-olds had a BMI in the normal weight range (n = 127, 72.6%). Twenty-six children were categorised as overweight or obese (14.8%). Girls were more likely to be categorised as underweight, overweight, and obese compared to boys.

CASE STUDY 3: Prevalence of anaemia

Overall, 47.1% (n = 124 out of 263) of children with at least one Hb reading were deemed to be anaemic at some point over the study period. We found 29.3% children were anaemic at 6 months of age, 43.8% at 12 months, 25.2% at 18 months, 28.1% at 2 years, 12.3% at 3 years, 7.3% at 4 years and 10.3% at 5 years of age (Table 2). The highest proportion of cases was observed at 12 months of age, affecting more girls than boys (n = 26, 24.7% vs n = 20, 19% respectively). Only four children in the whole sample were found to have severe anaemia (Hb <80g/L), all of which were recorded before the age of two. Longitudinal Hb results from 6 months to 5 years of age were only available for a single child. We did find a sample of 46 children with complete records from 6 to 18 months of age. Within this group, we identified 3 cases of chronic anaemia at 6 months that was still present at 12 months. Six out of the 9 cases identified at 12 months of age were new cases among those with normal Hb levels at 6 months. By 18 months of age, 5 of these new cases had recovered and were no longer considered anaemic.

Data quality assessment

Completeness

Completeness of key demographic data items for month and year of birth, sex and Aboriginal status was 100%. Completeness of all other variables differed greatly over time (Table 3). The most complete records were weight at 3 and 4 years of age (60.8% and 62.1% respectively). Birthweight was available for over half the cohort (54.2%). Haemoglobin test results were available for 32.9% of the cohort at 12 months of age and 34.1% of the cohort at 4 years of age. Low rates of completeness overall resulted in small sub-samples of children with serial measurements making longitudinal analysis difficult. Completion of ASQ-TRAK developmental screens in our cohort were low with only 15% having had at least one assessment. However, the number of ASQ-TRAK screenings had increased over time in line with the service's aspiration to screen eligible children from 2017 onwards.

Correctness

Seven of the extracted and analysed variables had some degree of inaccuracy (Table 3). The majority of discrepancies were caused by inconsistent values being entered across duplicate records for an individual client on the same day. Birth length had the highest number of conflicting values entered resulting in a correctness of 91.8%. Other minor data entry errors were considered random and mostly consisted of missing or extra

Table 2: Summary of haemoglobin test results for Congress children (n = 263)

	Total number HB tests performed at timepoint (%)	Missing (%)	Mean HB G/L (SD)	HB G/L range	Number children classified as anaemic at timepoint (%)
6 months	92 (28.8)	227 (71.2)	106.88 (SD 10.6)	73-130	27 (29.3)
12 months	105 (32.9)	214 (67.1)	110.10 (SD 11.6)	79-141	46 (43.8)
18 months	107 (33.5)	212 (66.5)	113.17 (SD 11.5)	67-138	27 (25.2)
2 years	96 (30.1)	223 (69.9)	113.48 (SD 10.7)	67-136	27 (28.1)
2.5 years	96 (30.1)	223 (69.9)	112.52 (SD 10.6)	82-133	29 (30.2)
3 years	89 (27.9)	230 (72.1)	120.00 (SD 11.4)	81-149	11 (12.3)
3.5 years	92 (28.8)	227 (71.2)	123.06 (SD 10.6)	94-146	9 (9.8)
4 years	109 (34.1)	210 (65.8)	125.55 (SD 11.4)	87-149	8 (7.3)
4.5 years	65 (20.4)	254 (79.6)	123.80 (SD 10.6)	88-140	5 (7.7)
5 years	39 (12.2)	280 (87.8)	123.10 (SD 10.5)	101-142	4 (10.3)

digits on recorded head circumference (97.2%), birth weight (97.7% correct), weight (98.5% correct) and height (99.8% correct) with very few outliers. Overall, Congress staff entered values into the clinical information system with a high degree of accuracy over the study period (average correctness of 99.2%).

Consistency

We found several key constructs recorded across multiple variables (refer to Appendix 5). Collection of the same (or very similar) types of information on separate forms within the clinical information system is likely due to different user requirements across service locations, departments, or teams. This has contributed to the majority of data disagreements. Where possible, we merged or appended identical constructs and harmonised similar constructs. Overall, the format or unit of values were all uniform (100% consistent). However, the variable labels and options for categorical data often differed and required harmonisation. These information discrepancies are not fatal, but they do require extensive data cleaning to correct. All of these issues impact the utility of existing EMRs for research purposes. A summary of data quality including completeness, correctness and consistency and subsequent recommendations is presented in Table 3.

Discussion

The present study aimed to assess the potential usability of Congress-held EMRs in longitudinal analysis of Aboriginal child health over the first five years of life. We evaluated the quality and completeness of 29 data items extracted for a sample of 373 young Aboriginal clients born in 2015 (Table 3). The findings indicate that the quality of the records was high, with an average correctness of 99.2% across all data items. The literature on quality improvement of EMRs suggests data errors can be reduced through techniques such as double entry, logic checks, and visual verification [32]. To further enhance data accuracy, spellchecking tools can also be employed to prevent typographical errors, and standardised coding can be achieved by using predictive drop-down lists that are mutually exclusive and cover all possible options [33]. Given the high quality (correctness) of the extracted data in this study, we

know that many of these approaches are already in place at Congress. All data items were recorded in the same format or unit indicating high consistency (100%).

The high proportion of missing or unavailable data poses the greatest challenge for those wishing to utilise Congress-held EMRs for longitudinal research. Certain data items proved to be more complete than others. According to Wells et al (2013), when data is missing it is difficult to determine whether the field was left empty due to a failure to ask, a failure to record or intentional omission by the service provider. 'Failure to document' is said to be especially prevalent when the value is negative. In such cases, fields are simply left blank [33]. This ambiguity makes it difficult to interpret change in health status over time. Missing data can also compromise the reliability of findings. Karahalios et al. (2012) state that missing data can reduce power and precision, although statistical methods exist to address this issue [34]. In our case studies, we used a complete-case analysis, which excludes participants with missing data. However, this method can introduce bias if the participants with missing data differ systematically from those with complete data. Other techniques, such as multiple imputation and Bayesian modelling may offer more robust alternatives when data are missing at random [34]. The preferred option, as noted by Wells et al (2013), would be to enhance completeness of data documentation from the outset [33]. To support this, actively involving clinicians and staff in the data quality assessments can help emphasise the value of maintaining data quality and completeness [35]. Congress could also consider automatic prompts to flag incomplete entries, particularly when recording negative values instead of leaving fields blank [33]. Mandating completion of HU5K templates and updating key socio-demographic fields at every check-up could also help. That said, we acknowledge the difficulty of asking staff, who are frequently under pressure to treat sick clients, to enter additional data during consultations. Congress staff have indicated that socio-demographic questions can be time-consuming and are sometimes perceived as intrusive by clients. While allocating additional time for data entry could improve data capture, this is unlikely to be feasible given the current service demands. Balancing the need for complete data with the practical burden on health service providers remains a critical challenge, which is being addressed through ongoing CQI efforts at Congress. [35]. Data linkage may offer a valuable solution

Table 3: Data quality assessment and recommendations

Data item	Completeness number (%)	Correctness (%)	Consistency of format/unit (%)	Suitability of data for longitudinal research and recommendations
Child month/year of birth	319 (100)	100	100	 Suitable for use in longitudinal analysis.
Child sex	319 (100)	100	100	 Suitable for use in longitudinal analysis.
Child Aboriginal status	319 (100)	100	100	 Suitable for use in longitudinal analysis.
Number of adults 	223 (69.9)	100	100	 Potential use in longitudinal analysis, should completeness increase. "Number of adults" captured on ad hoc basis &/or asked at multiple timepoints over study period. In lieu of another proxy measure for household crowding, recommend making "Number of adults" a mandatory field at each routine check-up to increase completeness which will facilitate longitudinal analysis.
Number of children 	219 (68.6)	100	100	 Potential use in longitudinal analysis, should completeness increase. "Number of children" captured on ad hoc basis &/or asked at multiple timepoints over study period. In lieu of another proxy measure for household crowding, recommend making "Number of children" a mandatory field at each routine check-up to increase completeness which will facilitate longitudinal analysis.
Anyone EVER smoked in house or car 	216 (67.7)	100	100	 Potential use in longitudinal analysis, should completeness increase. "Ever smoked" captured on ad hoc basis &/or asked at multiple timepoints over study period. Recommend making variable a mandatory field at each routine check-up to increase completeness which will facilitate longitudinal analysis.
Weight	Max (62.1) 4 yrs Min (19.7) 5 yrs	98.5	100	 Potential use in longitudinal analysis, should completeness increase. 59 data entry errors were found out of 3,924 weight measurements in 313 children. Some recoding required for conflicting duplicates recorded for same child on the same date. Lowest weight entered accepted as true. Recommend employing logic checks to enhance data entry accuracy. Facilitate inter and intra-department harmonisation of imputations of "weight" on same day by different health care providers/clinicians.
Height	Max (55.2) 4 yrs Min (12.5) 5 yrs	99.8	100	 Potential use in longitudinal analysis, should completeness increase. 3 data entry errors were found out of 1,880 height measurements in 290 children. Some recoding required for conflicting duplicates recorded for same child on the same date. Lowest height entered accepted as true. Recommend employing logic checks to enhance data entry accuracy. Facilitate inter and intra-department harmonisation of imputations of "height" on same day by different health care providers/clinicians.
Birth weight (kg)	173 (54.2)	97.7	100	 Potential use in longitudinal analysis, should completeness increase. Extensive data cleaning and recoding required for four children with conflicting duplicates records. Smallest birth weight entered accepted as true. Recommend employing logic checks to enhance data entry accuracy. Ideally "birth weight" should be collected retrospectively for all children at first encounter with new clients and set as a required field.
Gestational age (weeks)	160 (50.2)	97.5	100	 Potential use in longitudinal analysis, should completeness increase. Four children with conflicting duplicate records. Smallest gestational age (weeks) entered accepted as true. Some recoding required for conflicting duplicate records. Recommend verifying with estimated due date (EDD) recorded in mother's file or hospital discharge (if possible).
Birth length (cm)	146 (45.8)	91.8	100	 Potential use in longitudinal analysis, should completeness increase. 12 children with conflicting duplicate records. Smallest birth length entered accepted as true. Some recoding required for conflicting duplicate records. Recommend employing logic checks to enhance data entry accuracy. Ideally "birth length" should be collected retrospectively for all children at first encounter with new clients and set as a required field.
Head circumference (cm)	143 (44.8)	97.2	100	 Potential use in longitudinal analysis, should completeness increase. Some recoding required for four children with conflicting duplicate records. Smallest head circumference entered accepted as true. Recommend employing logic checks to enhance data entry accuracy.
5-minute APGAR score	142 (44.5)	100	100	 Potential use in longitudinal analysis, should completeness increase. Recommend verifying against mother's file or hospital discharge (if possible).
Immunisation status at 4 years of age	131 (41.1)	100	100	 Potential use in longitudinal analysis, should completeness increase. Could be verified with secondary source (such as the Australian Immunisation Register). Recommend making "Immunisation status" a mandatory field at each routine check-up to increase completeness which will facilitate longitudinal analysis.
Attendance at routine check ups	Max (40.1) 4 yrs Min (2.8) 5 yrs	100	100	 Potential use in longitudinal analysis, should completeness increase. Extensive data cleaning and recoding required to establish a derived "attendance" variable at each time point based on the presence or absence of any data recorded each of these key milestones (allowing +/- 30 days).

Continued

Table 3: Continued

Data item	Completeness number (%)	Correctness (%)	Consistency of format/unit (%)	Suitability of data for longitudinal research and recommendations
Attended ECE/Preschool EVER	123 (38.6)	100	100	 Potential use in longitudinal analysis, should completeness increase. Recommend further investigating how to best capture preschool or ECE attendance and dosage over time for children in the clinical information system. This would enable the service to longitudinally examine the impact on child's development and learning as they grow.
Housing type 	113 (35.4)	100	100	 Potential use in longitudinal analysis, should completeness increase. "Housing type" captured on ad hoc basis &/or asked at multiple timepoints over study period. Some recoding was required for overlapping free text "other" responses. Recommend making variable a mandatory field at each routine check-up to increase completeness via pre-coded or predictive text drop-down lists which will facilitate longitudinal analysis.
NT Hospital of birth	112 (35.1)	100	100	 Potential use in longitudinal analysis, should completeness increase, and minor modifications made. Some recoding was required for overlapping free text "other" responses. Recommend expanding data collection to include "place of birth" including non-hospital births and hospitals outside of the NT. Ideally collected retrospectively for all children, then prospectively from first encounter with new clients via pre-coded or predictive text drop-down lists.
Haemoglobin test results	Max (34.1) 4 yrs Min (12.2) 5 yrs	99.1%	100	 Potential use in longitudinal analysis, should completeness increase. 12 data entry errors were found out of 1,315 haemoglobin test results in 11 children. Errors found included conflicting answers recorded for same child on the same date. Lowest haemoglobin result entered accepted as true. Some recoding was required for conflicting duplicate records.
Main language spoken at home 	106 (33.3)	100	100	 Potential use in longitudinal analysis, should completeness increase, and minor modifications made. "Main language spoken at home" captured on ad hoc basis &/or asked at multiple timepoints over study period. Some recoding was required for overlapping free text "other" responses. Recommend expanding data collection to capture variable for all children at first encounter with health service via pre-coded or predictive text drop-down lists to facilitate longitudinal analysis.
Breastfeeding at hospital discharge	101 (31.7)	100	100	 Potential use in longitudinal analysis, should completeness increase. Some recoding/harmonisation with "feeding method" was required to separate out exclusive breastfeeding with mixed feeding. Recommend making this mandatory at first encounter with health service (maternal recall of breastfeeding duration has been shown to be valid and reliable, even after prolonged periods). This will enable the service to track exclusive breastfeeding rates longitudinally. Li R, Scanlon KS, Serdula MK. The validity and reliability of maternal recall of breastfeeding practice. Nutrition reviews. 2005 Apr 1;63(4):103-10.
Number of bedrooms 	98 (30.7)	100	100	 Potential use in longitudinal analysis, should completeness increase. "Number of bedrooms" captured on ad hoc basis &/or asked at multiple timepoints over study period. In lieu of another proxy measure for household crowding, recommend making "Number of bedrooms" a mandatory field at each routine check-up to increase completeness which will facilitate longitudinal analysis.
Number of languages spoken 	97 (30.4)	100	100	 Potential use in longitudinal analysis, should completeness increase, and minor modifications made. "Number of languages spoken at home" captured on ad hoc basis &/or asked at multiple timepoints over study period. Recommend expanding data collection to capture variable at first encounter with health service via pre-coded or predictive text drop-down lists to facilitate longitudinal analysis.
Main carer 	69 (21.6)	100	100	 Potential use in longitudinal analysis, should completeness increase, and minor modifications made. "Main carer" captured on ad hoc basis &/or asked at multiple timepoints over study period. Some recoding was required for overlapping free text "other" responses. Recommend expanding data collection to capture variable at every routine check-up via pre-coded or predictive text drop-down lists to facilitate longitudinal analysis.
Feeding method at 3 months	94 (29.5)	100	100	 Potential use in longitudinal analysis, should completeness increase. Recommend making "feeding method" a mandatory field at the 3, 6, 12 and 18-month routine check-up to enable the service to track exclusive breastfeeding rates and the introduction of solids longitudinally. Options must be mutually exclusive, ideally using pre-coded or predictive text drop-down lists.
ASQ-TRAK developmental screening	48 (15)	100	100	 Suitable for use in longitudinal analysis. Raw scores entered into the clinical information system from the ASQ-TRAK developmental screenings were all within plausible ranges. At the conclusion of each screen staff are required to manually enter the number of developmental domains with a score "below cut-off". We cross-checked the values entered by staff in each of the five domains, against this numeric value (number of domains below cut-off) and found them to be 100% correct.
Skin assessment	N/A	N/A	N/A	 Not suitable for longitudinal analysis. Partial data were supplied, but not confident of completeness. Coding of outcomes were inconsistent, as such could not confidently report data on skin assessments.
Recurrent/Persistent otitis media	N/A	N/A	N/A	 Not suitable for longitudinal analysis. Partial data were supplied, but not confident of completeness. Coding of outcomes were inconsistent, as such could not confidently report data on otitis media.

Continued

Table 3: Continued

Data item	Completeness number (%)	Correctness (%)	Consistency of format/unit (%)	Suitability of data for longitudinal research and recommendations
Respiratory assessment	N/A	N/A	N/A	 Not suitable for longitudinal analysis. Partial data were supplied, but not confident of completeness. Coding of outcomes were inconsistent, as such could not confidently report data on respiratory assessments.
Referrals made (including referrals for dental, hearing, and other further assessments)	N/A	N/A	N/A	 Not suitable for longitudinal analysis. Partial data were supplied, but not confident of completeness as "referrals made" is not a required field for each routine check-up or ad hoc encounter. This does not include referrals made following ASQ-TRAK developmental screenings.
Maternal age	N/A	N/A	N/A	 Requested, but not provided. Child record to be linked to Mother's record in the clinical information system.
Parity	N/A	N/A	N/A	 Requested, but not provided. Child record to be linked to Mother's record in the clinical information system.
Gravidity	N/A	N/A	N/A	 Requested, but not provided. Child record to be linked to Mother's record in the clinical information system.
Maternal complications	N/A	N/A	N/A	 Requested, but not provided. Child record to be linked to Mother's record in the clinical information system.
Mode of delivery	N/A	N/A	N/A	 Requested, but not provided. Child record to be linked to Mother's record in the clinical information system.
Hospital admissions	N/A	N/A	N/A	 Requested, but not provided. Few references to admissions in other fields but cannot report with confidence. Likely to be recorded in free-text clinical notes that we did not have access to.
Emergency Department admissions	N/A	N/A	N/A	 Requested, but not provided. Few references to admissions in other fields but cannot report with confidence. Likely to be recorded in free-text clinical notes that we did not have access to.

Completeness was calculated by dividing the number of available records (at least one record per child) by the total number of the cohort sample ($n = 319$). Correctness was calculated by dividing the number of records free from error by the total number of available records.



= suitable for use in longitudinal analysis.



= potential use in longitudinal analysis with some modifications.



= not suitable for longitudinal analysis without significant modifications or data linkage.



→ Data item collected ad hoc/randomly during non-routine or acute presentations.

for researchers needing to source missing socio-economic information. Ultimately, improving data completeness hinges on families accessing the service at key milestones—a more complex issue that will require further consideration.

In **Case Study 1** we found an average of 19.9 (ad hoc and routine) visits to the service per child over the 5-year study period. However, the number of visits varied greatly from 2 to 102. Non-routine visits to the service recorded in the clinical information system were to access a range of primary health care services and specialist programs aimed at promoting positive developmental, educational and health outcomes for children in the early years. Clinic location for each visit or service encounter was not specified in the dataset. Coding of

reasons for non-routine appointments and subsequent clinical outcomes were either inconsistent or not provided to us (i.e., only available in free-text fields). This limited our ability to confidently report visits by health outcomes or examine any location specific variation.

The official HU5K schedule includes well-child check-ups every six months from 2 to 4.5 years of age. However, we observed variation in the timing of service attendance relative to the scheduled milestones. This does not necessarily indicate disengagement but may instead reflect real-world patterns of service use for this population. Collapsing attendance into annual intervals would likely provide a more accurate reflection of participation, aligning more closely with the 40% reported

by Congress for the 2021-2022 period [36]. Attendance rates presented in Case Study 1 were calculated on the assumption that all children in the cohort had the opportunity to attend, though this does not account for variations in residency or service utilisation. Not all families engage with Congress, and some children come from remote communities outside of Alice Springs, which may explain some of the missing data. Therefore, the reported routine check-up attendance rates in Case Study 1 is likely to be an underestimation.

Over the study period, child and family services were delivered through decentralised clinics, which were responsible for routine well-child health checks. The data gathered at these decentralised clinics were included in our analysis. While the decentralised model facilitated continuity of care and chronic disease management, it resulted in the loss of a clear, central location for caregivers to bring well children for their routine check-ups. Moreover, the service had funding for three dedicated child health nurses to perform the routine check-ups across all clinics. However, recruiting and retaining staff for these roles over the study period proved challenging. In 2023, Congress decided to introduce a centralised program, aiming to promote routine attendance and address staffing challenges. Attendance rates in 2020 were likely impacted by COVID-19 induced access restrictions and workforce shortages. The reluctance of caregivers to attend appointments during 2020 likely played a role in the low attendance rates observed for 5-year-olds. Further analysis of an additional 6 months of follow-up data may reveal belated attendance at these appointments in early 2021.

Growth profiles of Congress children presented in **Case Study 2** were comparable to the WHO Standards until 12 months of age, after which Congress girls tended to weigh slightly more. From 3 years of age, our sample of Congress children measured below the WHO 50th percentile for height. This case study is based on a small subgroup of children with complete data. Given the extent of missing data, which is likely non-random, findings should be interpreted with caution. The point prevalence of overweight or obese observed in our cross-sectional sample of 4-year-old children is consistent with findings reported elsewhere [37]. Weight recorded at 4 years of age had the highest time specific degree of completeness at 62.1% and yet this only accounts for just over half of the cohort.

Anaemia in early childhood, commonly attributed to iron deficiency, can negatively impact growth and development [38]. Providing effective management of anaemia is a major area of focus for Congress. **Case Study 3** found about half of the cohort had at least one haemoglobin test result recorded between 6 months and 5 years of age. Anaemia peaked at 12 months of age (43.8%) which is comparable to rates reported previously by Congress in their 2016-2017 Annual Report for children living in one or two remote locations [39]. Similarly high rates (46% among 6- to 12-month-olds) were reported in a retrospective review of haemoglobin records of young children living in six remote Aboriginal communities across Northern Australia [40]. Data on anaemia among other groups of Australian children is limited, restricting our capacity to comparisons with the broader population [41]. Information regarding maternal anaemia during pregnancy, which is acknowledged as another significant risk factor for childhood anaemia, were not included

in this dataset. Clinicians at Congress are good at testing the most at-risk children including infants born to anaemic mothers, which may have skewed these findings. Nonetheless, the rate of anaemia reported in this analysis is still very high. The service is exploring ways to increase testing rates of all children across all sites. The current rate of anaemia among 6- to 12-month-olds tested across Congress clinics is currently 25%, reflecting a significant improvement since the end of the study period in 2020, when screening and testing for anaemia were notably affected by the COVID-19 pandemic [42].

Omission of several important domains greatly compromised the utility of the extracted dataset. A comprehensive list of data items (or their proxy measures) frequently included in the longitudinal analysis of Indigenous children worldwide [21] were requested of the service. However, critical measures, including exposure, outcome and various covariates of interest relating to the child's physical, social and emotional wellbeing were not available (see Appendix 1). Furthermore, key data on the mother's health during pregnancy and socio-demographic factors (such as parental age, education, occupation, income, and place of residence) were either missing, not provided, or stored in free-text fields, making them unavailable to the research team as anonymity could not be easily assured. Access to these free-text fields could have provided valuable context, captured patient or caregiver concerns that aren't easily categorised, and revealed patterns or emerging issues not reflected in predefined categories. The absence of this information greatly impacted our ability to explore associations between exposures and health outcomes or examine group differences.

The clinical information system currently lacks the capability to link mothers' records with their children's. The service is looking into how this functionality might be implemented, which would be a valuable improvement. This enhancement will help address some of the data accessibility issues encountered in this analysis. Obtaining consent to link EMRs to other external data sources (e.g. Alice Springs Hospital) may be another way to address these data gaps in the future. Artificial intelligence and machine learning technologies, such as natural language processing tools, have emerged as a means to assist in the extraction of relevant data from free-text clinical notes [15]. Likewise, detection algorithms can now locate and redact personal identifiers to protect patient confidentiality [43]. To bridge the data gap, future work might involve exploring these options too.

Data used for this analysis were extracted from a single database for clinics and services located in and around the township of Alice Springs. Data on children accessing the remote clinics or services were not available at the time of the initial extraction. Since then, Congress has consolidated all of its Communicare data into a single database. This limitation may account for some of the missing or incomplete data, particularly for transient cohort members who moved in and out of town during the study period. Another limitation of our data quality assessment is its reliance on internal validation or cross-checking within the dataset. While this approach has been used elsewhere [44], the gold standard is to compare EMRs to manual review of paper records or files [45]. This was beyond our capacity and outside the scope of our approval granted by the Congress research sub-committee. The conclusions drawn in our study are also

derived from historical data. It is important to note that additional quality assurance checks and improvements may have been implemented since.

Despite these limitations, we have identified several barriers in our example analyses and identified what is required to successfully harness Congress-held EMRs for future longitudinal research. Further exploration of how completeness can be improved to achieve this goal is required. This study was intended to be practical, relevant, and directly useful to the service in their endeavour to harness much needed longitudinal data on the health and wellbeing of local Aboriginal children. Our case studies have deepened our understanding of which additional data is required to facilitate effective longitudinal research using existing EMRs. These learnings may be of use to other Aboriginal health services across Australia, many of whom use the same or similar clinical information systems. We are aware of several other studies that have leveraged EMRs to facilitate research [46–49]. Some studies have even evaluated the quality of EMRs compared to paper files [50, 51] but none have specifically focused on using EMRs in longitudinal research of young Aboriginal children. A natural progression of this work would be to compare our findings with data extracted directly from client files (where available). Another possible extension would be to investigate the validity of using multiple imputation to handle missing values.

Conclusion

Analysis of a series of key data items extracted from EMRs pertaining to young Aboriginal clients of Congress (born in 2015) found that missing or incomplete data, or data in an inaccessible free-text format were the greatest challenges. Overall, data entered into the clinical information system by staff were of high quality and consistency. Future use of Congress-held EMRs for longitudinal research will be contingent on increasing rates of data capture at regular intervals in early life. The findings presented here will be useful in developing a set of principles and common data standards for 'research-ready' longitudinal datasets at Congress. This would pave the way to link in-house early life longitudinal data (with dynamic consent) to prospectively collected cohort data, leading to new insights and discoveries that ultimately improve patient care.

Acknowledgments

Aboriginal author (SE) collaborated closely with non-Aboriginal authors (CLJ, JB, VB, RW, SGuo, AD and SG) to produce this article. CLJ conceived the initial design of the study and organised the data request. VB facilitated the data extraction with assistance from the Congress Data Manager and Congress CQI staff. CLJ and SGuo conducted the data preparation. CLJ completed the analysis with assistance from SGuo. CLJ wrote the first draft of the manuscript. SG, AD, and SE advised on research conduct from inception to completion. SG, JB, VB, RW, SE and AD appraised the analysis process conducted by CLJ and SGuo. All authors read, revised and approved the final manuscript. The authors alone are responsible for the views expressed in this article and they

do not necessarily represent the views, decisions, or policies of the institutions with which they are affiliated.

Statement of conflicts of interest

The authors declare that they have no conflicts of interest with respect to the research, authorship and/or publication of this article.

Ethics statement

The study was endorsed by the Central Australia Aboriginal Congress Research Sub-Committee on 5 February 2021, prior to receiving formal ethics approval from the Royal Children's Hospital Human Research Ethics Committee (2019.155) on September 13, 2021, and the Central Australian Human Research Ethics Committee (CA-19-3519) on September 24, 2021. An amendment request to include additional data items was approved by the Congress Research Manager on April 14, 2022, and the Royal Children's Hospital Human Ethics Research Committee (2019.155) on 6 May 2022, and the Central Australian Human Research Ethics Committee (CA-19-3519) on 19 May 2022. The Congress Research Sub-Committee includes Aboriginal leadership and exists to support and promote research that reflects and is responsive to the needs of the local Aboriginal community. This study has been conducted in accordance with the ethical guidelines for best practice in research involving Aboriginal communities [52–59]. Our analysis has been written from a strengths-based approach and in partnership with Congress as co-contributors and co-authors. We report this study in accordance with the RECORD statement for observational studies using routinely collected data (Appendix 6).

Data availability statement

Data are available from the authors upon reasonable request and with permission of the Central Australian Aboriginal Congress Research Sub-Committee via the corresponding author.

Funding

CLJ has been supported by an Australian Government Research Training Program (RTP) Scholarship through the University of Melbourne. John T Reid Charitable Trusts provided philanthropic support for this research. All opinions expressed in this paper are those of the authors alone.

References

1. McEwen BS. Early life influences on life-long patterns of behavior and health. Mental retardation and developmental disabilities research reviews. 2003;9(3):149-54. <https://doi.org/10.1002/mrdd.10074>
2. Kuo AA. Life Course Health Development. In: Kuo AA, Collier RJ, Stewart-Brown S, Blair M, editors. Child

- Health: A Population Perspective: Oxford University Press; 2015. p. 111-132.
3. Shonkoff JP, Slopen N, Williams DR. Early childhood adversity, toxic stress, and the impacts of racism on the foundations of health. *Annual Review of Public Health*. 2021;42:115-34. <https://doi.org/10.1146/annurev-publhealth-090419-101940>
4. Kalmakis KA, Chandler GE. Health consequences of adverse childhood experiences: A systematic review. *Journal of the American Association of Nurse Practitioners*. 2015;27(8):457-65. <https://doi.org/10.1002/2327-6924.12215>
5. Australian Institute of Health and Welfare (AIHW). Determinants of health for First Nations people 2024 [cited 2024 Dec 6]. Available from: <https://www.aihw.gov.au/reports/australias-health/social-determinants-and-indigenous-health>.
6. Bond CJ, Singh D. More than a refresh required for closing the gap of Indigenous health inequality. *Medical Journal of Australia*. 2020;212(5):198-9. e1. <https://doi.org/10.5694/mja2.50498>
7. Griffiths KE. Visibility, power and equity: Using Aboriginal and Torres Strait Islander data to drive equity in Australia. *Journal and Proceedings of the Royal Society of New South Wales*; 2023: Royal Society of New South Wales Sydney. <https://search.informit.org/doi/abs/10.3316/informit.201951421713813>
8. Angell B, Eades S, Jan S. To Close the Gap we need to identify the best (and worst) buys in Indigenous health. *Australian and New Zealand Journal of Public Health*. 2017;41(3):224-6. <https://doi.org/10.1111/1753-6405.12612>
9. Lloyd-Johnsen C, Eades S, D'Aprano A, Goldfeld S. What's data got to do with it? A scoping review of data used as evidence in policies promoting the health of Aboriginal and Torres Strait Islander children in the Northern Territory, Australia. *Health Promotion Journal of Australia*. 2023;34(2):443-71. <https://doi.org/10.1002/hpja.670>
10. Walter M. The voice of Indigenous data: Beyond the markers of disadvantage. *Griffith Review*. 2018(60):256-63. <https://search.informit.org/doi/10.3316/informit.586241932732209>
11. Kukutai T, Taylor J. Indigenous data sovereignty: Toward an agenda: ANU press; 2016. <http://www.jstor.org/stable/j.ctt1q1crgf>
12. Central Australian Aboriginal Congress (Congress). Central Australian Aboriginal Congress Board Communique Issue 6, February 2014 Alice Springs: Congress; 2014 [cited 2023 Dec, 12]. Available from: https://www.caac.org.au/wp-content/uploads/pdfs/07-2014-February_Board-Communique_FINAL.pdf,
13. Barrett D, Noble H. What are cohort studies? *Evidence-Based Nursing*. 2019. <https://doi.org/10.1136/ebnurs-2019-103183>
14. Caruana EJ, Roman M, Hernández-Sánchez J, Solli P. Longitudinal studies. *Journal of Thoracic Disease*. 2015;7(11):E537. <https://doi.org/10.3978/j.issn.2072-1439.2015.10.63>
15. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. *Annual Review of Public Health*. 2016;37:61-81. <https://doi.org/10.1146/annurev-publhealth-032315-021353>
16. Jorm L. Routinely collected data as a strategic resource: priorities for methods and workforce. *Public Health Research & Practice*. 2015. <https://doi.org/10.17061/phrp2541540>
17. Tingay KS, Bandyopadhyay A, Griffiths L, Akbari A, Brophy S, Bedford H, et al. Record linkage to enhance consented cohort and routinely collected health data from a UK birth cohort. *International Journal of Population Data Science*. 2019;4(1). <https://doi.org/10.23889/ijpds.v4i1.579>
18. Calderwood L, Lessof C. Enhancing longitudinal surveys by linking to administrative data. *Methodology of Longitudinal Surveys*. 2009:55-72. <https://doi.org/10.1002/9780470743874.ch4>
19. Thuraingam S, Chondros P, Dowsey MM, Spelman T, Garies S, Choong PF, et al. Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment. *BMC Medical Informatics and Decision Making*. 2021;21(1):1-11. <https://doi.org/10.1186/s12911-021-01669-6>
20. Li L, O'Neil L. Mothers and babies 2015: Northern Territory Midwives' Collection. Darwin: Department of Health; 2018. [cited 2024 Dec, 6]. Available from: <https://digitallibrary.health.nt.gov.au/nthealthserver/api/core/bitstreams/267c742dae47-4e93-b471-67ab16c24521/content>.
21. Lloyd-Johnsen C, Eades S, McNamara B, D'Aprano A, Goldfeld S. A global perspective of Indigenous child health research: a systematic review of longitudinal studies. *International Journal of Epidemiology*. 2021;50(5):1554-68. <https://doi.org/10.1093/ije/dyab074>
22. Liaw S-T, Taggart J, Dennis S, Yeo A, editors. Data quality and fitness for purpose of routinely collected data—a general practice case study from an electronic practice-based research network (ePBRN). *AMIA Annual Symposium Proceedings*; 2011: American Medical Informatics Association.
23. Kruske S, Donovan J, McKinnon T, Nuttall L. Healthy Under 5 Kids Program: education package: DHF

- and CDU; 2009. [cited 2024 Dec, 6]. Available from: <https://digitalibrary.health.nt.gov.au/entities/publication/9e7424de-dc36-48a0-a2c5-f89fddbe2eda>.
24. World Health Organization (WHO). WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. Geneva, Switzerland: World Health Organization; 2006. [cited 2024 Dec, 6]. Available from: <https://www.who.int/publications/i/item/924154693X>.
25. Remote Primary Health Care Manuals. CARPA Standard Treatment Manual (8th Edition). Alice Springs, Northern Territory: Flinders University; 2022. [cited 2024 Dec, 6]. Available from: <https://remotephcmmanuals.com.au/document/35499.html>.
26. Australian Institute of Health and Welfare (AIHW). Australia's mothers and babies 2015—in brief Canberra 2017 [cited 2024 Dec, 6]. Available from: <https://www.aihw.gov.au/reports/mothers-babies/australias-mothers-babies-2015-inbrief/summary>.
27. Brown S, Stuart-Butler D, Leane C, Glover K, Mitchell A, Deverix J, et al. Initiation and duration of breastfeeding of Aboriginal infants in South Australia. *Women and Birth*. 2019;32(3):e315-e22. <https://doi.org/10.1016/j.wombi.2018.07.009>
28. Simpson S, Eadie T, Khoo ST, Titmuss A, Maple-Brown LJ, Thompson R, et al. The ASQ-TRAK: validating a culturally adapted developmental screening tool for Australian Aboriginal children. *Early Human Development*. 2021;163:105481. <https://doi.org/10.1016/j.earlhumdev.2016.08.005>
29. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ*. 2000;320(7244):1240. <https://doi.org/10.1136/bmj.320.7244.1240>
30. Cole TJ, Flegal KM, Nicholls D, Jackson AA. Body mass index cut offs to define thinness in children and adolescents: international survey. *BMJ*. 2007;335(7612):194. <https://doi.org/10.1136/bmj.39238.399444.55>
31. Cole TJ, Lobstein T. Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatric obesity*. 2012;7(4):284-94. <https://doi.org/10.1111/j.2047-6310.2012.00064.x>
32. Amoakoh-Coleman M, Kayode GA, Brown-Davies C, Agyepong IA, Grobbee DE, Klipstein-Grobusch K, et al. Completeness and accuracy of data transfer of routine maternal health services data in the greater Accra region. *BMC Research Notes*. 2015;8:1-9. <https://doi.org/10.1186/s13104-015-1058-3>
33. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *Egms*. 2013;1(3). <https://doi.org/~10.13063/2327-9214.1035>
34. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC Medical Research Methodology*. 2012;12(1):1-10. <https://doi.org/10.1186/1471-2288-12-96>
35. Canaway R, Boyle D, Manski-Nankervis J-A, Gray K. Identifying primary care datasets and perspectives on their secondary use: a survey of Australian data users and custodians. *BMC Medical Informatics and Decision Making*. 2022;22(1):94. <https://doi.org/10.1186/s12911-022-01830-9>
36. Central Australian Aboriginal Congress (Congress). Congress 2021-2022 Annual Report Alice Springs, Northern Territory 2022 [cited 2024 Dec, 6]. Available from: <https://www.caac.org.au/wp-content/uploads/2022/12/Congress-2022-Annual-Report.pdf>.
37. Thurber KA, Dobbins T, Neeman T, Banwell C, Banks E. Body mass index trajectories of Indigenous Australian children and relation to screen time, diet, and demographic factors. *Obesity*. 2017;25(4):747-56. <https://doi.org/10.1002/oby.21783>
38. Leonard D, Buettner P, Thompson F, Makrides M, McDermott R. Early childhood anaemia more than doubles the risk of developmental vulnerability at school-age among Aboriginal and Torres Strait Islander children of remote Far North Queensland: Findings of a retrospective cohort study. *Nutrition & Dietetics*. 2020;77(3):298-309. <https://doi.org/10.1111/1747-0080.12602>
39. Central Australian Aboriginal Congress (Congress). Congress 2016-2017 Annual Report 2017 [cited 2024 Dec, 6]. Available from: https://www.caac.org.au/wp-content/uploads/pdfs/2016-17_Congress_Annual_Report_web.pdf.
40. Aquino D, Leonard D, Hadgraft N, Marley JV. High prevalence of early onset anaemia amongst Aboriginal and Torres Strait Islander infants in remote northern Australia. *Australian Journal of Rural Health*. 2018;26(4):245-50. <https://doi.org/10.1111/ajr.12403>
41. Leonard D, Buttner P, Thompson F, Makrides M, McDermott R. Anaemia in early childhood among Aboriginal and Torres Strait Islander children of Far North Queensland: a retrospective cohort study. *Australian and New Zealand Journal of Public Health*. 2019;43(4):319-27. <https://doi.org/10.1111/1753-6405.12911>
42. Boffa J. Current rate of anaemia among 6- to 12-month-olds. Personal Correspondence to Lloyd-Johnsen, C (cat.lloydjohnsen@mcri.edu.au) on Sept 19, 2024. [cited 2024 Dec, 6]

43. Garies S, Cummings M, Forst B, McBrien K, Soos B, Taylor M, et al. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction, and processing in Alberta. *International Journal of Population Data Science*. 2019;4(2). <https://doi.org/10.23889/ijpds.v4i2.1132>
44. Woods JA, Johnson CE, Allingham SF, Ngo HT, Katzenellenbogen JM, Thompson SC. Collaborative data familiarisation and quality assessment: reflections from use of a national dataset to investigate palliative care for Indigenous Australians. *Health Information Management Journal*. 2021;50(1-2):64-75. <https://doi.org/10.1177/1833358320908957>
45. Young JC, Conover MM, Jonsson Funk M. Measurement error and misclassification in electronic medical records: methods to mitigate bias. *Current Epidemiology Reports*. 2018;5:343-56. <https://doi.org/10.1007/s40471-018-0164-x>
46. Hoy WE, Swanson CE, Hope A, Smith J, Masters C. Evidence for improved patient management through electronic patient records at a Central Australian Aboriginal Health Service. *Australian and New Zealand Journal of Public Health*. 2014;38(2):154-9. <https://doi.org/10.1111/1753-6405.12195>
47. Dzidowska M, Lee KK, Conigrave JH, Dobbins TA, Hummerston B, Wilson S, et al. Support for Aboriginal health services in reducing harms from alcohol: 2-year service provision outcomes in a cluster randomized trial. *Addiction*. 2022;117(3):796-803. <https://doi.org/10.1111/add.15712>
48. Conigrave JH, Harrison KH, Lee KK, Dobbins TA, Hummerston B, Hayman N, et al. Support can increase use of the AUDIT-C in Australian Aboriginal Community controlled health services: A cluster randomized trial. *Addiction*. 2021;116(9):2304-15. <https://doi.org/10.1111/add.15428>
49. Regan T, Paul C, Ishiguchi P, D'Este C, Koller C, Forshaw K, et al. Comparison of Two Sources of Clinical Audit Data to Assess the Delivery of Diabetes Care in Aboriginal Communities. *International Journal of Environmental Research and Public Health*. 2017;14(10):1236. <https://doi.org/10.3390/ijerph14101236>
50. Davis S, Reeve C, Humphreys JS. How good are routinely collected primary healthcare data for evaluating the effectiveness of health service provision in a remote Aboriginal community? *Rural and Remote Health*. 2015;15(4):302-10. <https://doi.org/10.22605/RRH2804>
51. Hill I, Johnson D, Scrimgeour D, McDermott R. Accuracy of national key performance indicator reporting from two Aboriginal medical services: potential to underestimate the performance of primary health care. *Australian Health Review*. 2017;42(4):453-60. <https://doi.org/10.1071/AH16251>
52. Central Australian Aboriginal Congress (Congress). A guide for health researchers working in Central Australia Alice Springs: Central Australian Aboriginal Congress; 2021 [cited 2024 Dec, 6]. Available from: <https://www.caac.org.au/wp-content/uploads/2022/05/Doing-It-Right-web-compressed.pdf>.
53. NHMRC. Values and Ethics: Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Health Research. Canberra: National Health and Medical Research Council (NHMRC); 2003.
54. NHMRC. Keeping research on track: a guide for Aboriginal and Torres Strait Islander peoples about health research ethics. National Health & Medical Research Council; 2006. Report No.: 1864962976.
55. NHMRC. National statement on ethical conduct in human research 2007 (updated 2018). Commonwealth of Australia Canberra; 2007.
56. NHMRC. Keeping research on track II: A companion document to ethical conduct in research with Aboriginal and Torres Strait Islander Peoples and communities: Guidelines for researchers and stakeholders. National Health & Medical Research Council; 2018.
57. NHMRC. Australian code for the responsible conduct of research: National Health and Medical Research Council; 2018.
58. NHMRC. Ethical conduct in research with Aboriginal and Torres Strait Islander peoples and communities: guidelines for researchers and stakeholders: National Health and Medical Research Council; 2018.
59. AIATSIS. AIATSIS Code of Ethics for Aboriginal and Torres Strait Islander Research. 2020.



Appendix 1: Study 3 – requested data items

study variables	
Category	Data items/domains
Family health, history and demographics	– Month and year of birth (M/YYYY) – Gestational age at birth – Postcode/suburb at time of each visit (if known) – Language spoken at home – First language (dropdown list) – Main language spoken at home (dropdown list) – Number of languages spoken (total) – Interpreter required (Yes/No) – Type of housing (public, private, town camp) – Household/Family income (if known) – Financial stress (tick box) – Lack of social support/isolation (tick box) – Family worries or concerns (Yes/No collected HU5K child checks) – Recent family stress (tick box) – Family/household food (in)security (if recorded) – Parent/carer education (including highest level of school completed) – Parent/carer marital status – Parent/carer employment status – Family/Carer situation (dropdown list) – Housing conditions (including any urgent housing repairs needed) – Number of children living in the house – Number of siblings – Number of adults living in the house – Number of bedrooms in house – Mother or carer with disability – Other children in family with disability – Siblings with malnutrition or FTT – Family barriers – total (numeric) – Family Hx of respiratory problems (Yes/No) – Family Hx of ear problems (Yes/No) – Family Hx of eye problems (Yes/No) – Parent/Carer mental health issues (yes/no) – Substance abuse in immediate family (Yes/No) – Substance use history taken/updated (tick box) – Parent/Carer mental health issues (yes/no)
Early childhood Education	– Childcare/ Preschool Readiness Program (PRP)/ Preschool enrolment data ○ Days enrolled ○ Months enrolled
Education	– Timestamped records of 3A intensive sessions attended – Timestamped record of Learning and Cultural plans in place – Any recorded reference to primary school performance (e.g., Grade 3 or 5 NAPLAN results, Australian Early Development Census (AEDC) results) – Educational progress (notes)

Continued

Appendix 1: Study 3 – requested data items

study variables	
Category	Data items/domains
Health and development	– Birth details including: Mode of birth, duration of labour (numeric), premature birth <37 weeks (tick box), Twin or triplets? Yes/No, APGAR scores, IUGR/Premature/Low Birth Weight (Yes/No), list any birth/neonatal/abnormalities, list of maternal complications, feeding method at birth, history of alcohol or drug use in pregnancy, NT Hospital of birth or place of birth, first time or young mother, parity/gravidity, weeks antenatal care provided, EPDS score (if known). – Baby/Child's Gender – Indigenous status (child's father, child's mother) – Birth weight (grams) – Weight (timestamped) – Height or length including growth centiles (timestamped) – Feeding method (at birth and serial timepoints) – Breastfeeding (exclusively, last 24 hours, at discharge hospital) – Introduction of solids – Diet assessments – Major diagnoses (such as, anaemia, asthma, ear infections etc.) – Pathology results (if known) – Immunisation records (vaccine type/dose/date administered) – Timestamped data/scores and notes for developmental screening/checks (such as ASQ-TRAK) – Timestamped data/scores and notes for developmental assessments, including Paediatric assessments, and formal developmental assessments (such as, The Bayley Scales of Infant and Toddler Development etc.) – Timestamped data/scores and notes for any language assessments – Timestamped data/scores and notes for any educational assessments – Is child exposed to passive smoking? – Does anyone smoke in the house or car? – Developmental flags (drop down list) – Does your child sleep well at night? (Yes/No) – Development: Toilet training (nappies) – Child barriers total (numeric) – Child mood assessment (if recorded) – Parent/child interaction (MEMO) – Medication
Healthcare and service utilisation	– visits to Congress clinics and services (including location of clinic or service, date of visit) – all-cause hospital admissions (including date of admission, length of stay, reason for admission, ICD-10 codes etc.) – all-cause ER presentations (including date of presentation, reason for presentation, ICD-10 codes etc.) – referrals to allied health providers/ CYATS/other specialist services and date/outcome if known – visits to allied health providers/ CYATS/other specialist services and date/outcome if known – Interagency / partner meetings held and date/outcome if known – Utilisation of Congress transport – Child identified at risk (Yes/No) – Child Protection notification made (Yes/No) – Siblings notified to FACS (Yes/No) – Social worker (BoT) assessments (such as scores from Strengths and Difficulties Questionnaire, Strong Souls Assessment Tool, Wechsler – Non-Verbal scale of ability, Wechsler Individual Achievement Test, Adaptive Behavior Assessment System)

Appendix 2: Study 3 - Extracted data items

Variable name	Birth	~8 Weeks	4 Months	6 Months	9 Months	12 Months	18 Months	2 Years	2.5 Years	3 Years	3.5 Years	4 Years	4.5 Years	5 Years
Child month/year of birth	☑													
Child sex	☑													
Child Aboriginal status	☑													
Birth weight (kg)	☑													
Birth length (cm)	☑													
Gestational age (weeks)	☑													
NT hospital of birth	☑													
5-minute APGAR score	☑													
Head circumference (cm)	☑													
Breastfeeding at hospital discharge	☑													
Weight (kg)		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Height (cm)		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
BMI		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Feeding method		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Skin assessment		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Recurrent/ Persistent otitis media				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Respiratory assessment				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Dental referral				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Hearing referral				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Other referrals	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
Immunisation status			☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Haemoglobin check				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Iron deficiency treatment	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
Main carer/ Family carer situation*	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
Does anyone smoke in the house or car?		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
How many adults live in your house?		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
How many children live in your house?		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Housing type	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
How many bedrooms are in the house?		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
How many languages do you speak?	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
First language	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
Main language spoken at home	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔	↔
ASQ-TRAK raw score (Domain 1) communication				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
ASQ-TRAK raw score (Domain 2) gross motor				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
ASQ-TRAK raw score (Domain 3) fine motor				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
ASQ-TRAK raw score (Domain 4) problem solving				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
ASQ-TRAK raw score (Domain 5) personal-social				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Number of domains below cut-off				☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Referrals made following ASQ-TRAK assessments				↔		↔	↔	↔		↔		↔		
Has the child started preschool? *												☑		

☑ Data items collected as part of routine healthcare.

↔ Data item collected ad hoc/randomly during non-routine or acute presentations.

*Derived composite variable generated by merging 1 or more matched/harmonised variables.



Appendix 3. Study 3 - Detailed data preparation steps

In preparing the dataset for analysis, we undertook the following steps:

- (1) Records of ineligible children were removed (i.e., children not born in 2015, and/or with only a single visit recorded over the study period).
- (2) Duplicate records (i.e., the same variable populated on the same clinical date for the same child) were documented. We cross-checked the values in the duplicate records to see if they were consistent. If continuous values that were found to be inconsistent, we took the smallest value to true. Duplicate records were then dropped.
- (3) Each variable was checked for outliers or implausible values. After documenting the number of outliers or implausible values, we re-coded these as missing.
- (4) We assigned the first day of each month as a proxy “day” for their “date of birth” with Communicare provided “month/year” variable.
- (5) Then we derived a new variable “child age at visit” using new “date of birth” variable and each clinical

date. “Child age at visit” was created to categorise each visit into age-based health assessment milestones. The child's age at each visit (in years) was rounded to the nearest 0.25 using the STATA command `gen cage = round(clinicalitemage, 0.25)`, enabling visits to be grouped within three-month windows of scheduled milestone ages. For example, if a child attended a visit at 12.2 months, this would be grouped under the 12-month milestone, as it falls within the 10.5 to 13.5-month window. To better reflect observed patterns of attendance, we combined the First Visit and 8-week visit into a single milestone (0–8 weeks).

- (6) Checked derived age at visit for duplicate records (i.e., the same variable populated on the same clinical date for the same child). We cross-checked the values in the duplicate records to see if they were consistent. If continuous values that were found to be inconsistent, we took the smallest value to true. Duplicate records were then dropped.
- (7) We then reshaped the raw data into a format conducive to statistical analysis with 1 row per child and each child's associated data in columns to the right.

The earliest clinical date per child was considered to be their baseline episode of care date. Multiple records occurring on the same day were defined as one visit.



Appendix 4: Study 3 – Mean, median weight/height/bmi

Table a: Weight, height and BMI in sub-sample of Congress boys (n=9) with complete longitudinal data collected at birth, 6, 12, 18 months, 2, 3 and 4 years of age

Congress boys (n = 9)															
Weight (kg)						Height/Length (cm)					BMI (kg/m ²)				
	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
Birth	3.18	0.48	3.23	2.20	3.89	50.20	2.39	51.00	47.00	53.50	-	-	-	-	-
6 months	8.11	1.34	7.94	5.49	9.94	66.33	3.07	67.00	59.00	70.50	-	-	-	-	-
12 months	10.10	1.50	10.18	7.50	12.10	75.57	2.61	76.00	71.50	79.50	-	-	-	-	-
18 months	11.52	2.20	10.85	8.61	16.22	81.77	3.37	82.50	77.00	86.00	-	-	-	-	-
2 years	12.96	2.57	12.70	9.96	18.90	85.83	3.53	88.00	79.00	89.00	-	-	-	-	-
3 years	14.16	1.94	14.00	11.60	18.20	92.67	4.39	93.70	84.00	100.00	16.51	2.19	16.41	13.54	21.50
4 years	16.37	2.57	15.9	12.70	21.40	100.11	5.69	101.50	88.00	106.00	16.29	1.66	15.88	13.36	19.59

Table b: Weight, height and BMI in sub-sample of Congress girls (n=9) with complete longitudinal data collected at birth, 6, 12, 18 months, 2, 3 and 4 years of age

Congress girls (n = 9)															
Weight (kg)						Height/Length (cm)					BMI (kg/m ²)				
	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
Birth	2.99	0.39	2.98	2.53	3.70	50.27	1.69	50.00	48.00	53.50	-	-	-	-	-
6 months	7.23	0.95	7.40	5.75	8.74	65.33	3.36	65.00	61.00	71.00	-	-	-	-	-
12 months	9.26	0.89	9.28	7.57	10.92	74.33	3.67	75.00	66.00	78.00	-	-	-	-	-
18 months	10.50	1.17	10.78	8.53	12.10	79.77	3.11	79.00	75.50	84.00	-	-	-	-	-
2 years	12.07	1.42	12.10	9.10	13.70	84.38	4.02	86.00	75.50	88.00	-	-	-	-	-
3 years	14.63	1.75	14.50	11.60	17.30	93.46	4.41	93.20	86.00	99.30	16.73	1.50	16.63	14.87	20.00
4 years	16.78	1.91	17.00	13.40	19.50	101.11	5.30	101.00	92.00	107.00	16.40	1.31	15.83	14.65	18.42



Appendix 5: Study 3 - Constructs captured across multiple data items

Exact match	Partial match	No match
– Birthweight (SOURCE: Extraction identifier) – Birthweight (kg) SOURCE: (Birth details form) – Baby's birthweight (kg) SOURCE: (low birth weight/premature infant plan) – Gestational age at birth (weeks) SOURCE: (Birth details form) – Gestational age at birth (weeks) SOURCE: (low birth weight/premature infant plan) – Gestational age at birth (weeks) SOURCE: (ANFPP infant health care form) – Number of adults per household (Numeric, discrete) SOURCE: (HU5K child health check) – How many adults live in your house? (Numeric, discrete) SOURCE: (Social and Family History) – Number of children per household (Numeric, discrete) SOURCE: (HU5K child health check) – How many children live in your house? (Numeric, discrete) SOURCE: (HU5K child health check)	– Breastfeeding exclusively Yes/No SOURCE: (HU5K; Child Hx and Risk Assessment) – Breastfeeding at hospital discharge SOURCE: (HU5K; Child Hx and Risk Assessment) – Feeding method (drop down list – “Breast only”) SOURCE: (Birth details form) – Feeding method (drop down list – “Breast only”) SOURCE: (HU5K child health check) – Has the child started preschool? – Enrolled at preschool? – Enrolled at CHAD – Date enrolled in preschool	– First language (drop down list, plus “other” option for free text) SOURCE: (PRP initial assessment) – Main language spoke at home (drop down list, plus “other” option for free text) SOURCE: (Social and Family History) – How many languages do you speak? (Numeric, discrete) SOURCE: (Social and Family History)



Appendix 6. Study 3 – Record statement

The REporting of studies Conducted using Observational Routinely collected health Data (RECORD) statement is a checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

Source: Benchimol EI et al. (2015) The REporting of studies Conducted using Observational Routinely collected health Data (RECORD) Statement. PLoS Med 12(10): e1001885. <https://doi.org/10.1371/journal.pmed.1001885>.

	Item no.	STROBE items	Location in manuscript where items are reported	STROBE items	Location in manuscript where items are reported
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found		RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included. RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract. RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Title page
Introduction Background/rationale	2	Explain the scientific background and rationale for the investigation being reported			(Introduction) Pages 3-4
Objectives	3	State specific objectives, including any prespecified hypotheses			(Introduction) Page 4
Methods Study design	4	Present key elements of study design early in the paper			(Methods) Pages 4-6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection			(Abstract Introduction and Methods) Pages 2-6
Participants	6	(a) Cohort study Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up		RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided. RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided. RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	(Methods) Pages 4-6 Supplementary Appendix 3

	Item no.	STROBE items	Location in manuscript where items are reported	STROBE items	Location in manuscript where items are reported
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.		RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	(Methods) Pages 4-6 Supplementary Appendix 2 and Table 3 (pages 14-18)
Data sources/measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group			(Methods) Pages 4-6 Supplementary Appendices 2, 3 and 5
Bias	9	Describe any efforts to address potential sources of bias			Bias/Limitations discussed on Pages 19-21
Study size	10	Explain how the study size was arrived at			(Methods) Pages 4-6 Supplementary Appendix 3
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why			(Methods) Pages 4-6 Supplementary Appendix 3
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) Cohort study - If applicable, explain how loss to follow-up was addressed			(Methods) Pages 4-6 Supplementary Appendix 3
Data access and cleaning methods				RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	(Methods) Pages 4-6 Supplementary Appendix 3
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the study (e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed) (b) Give reasons for non-participation at each stage. (c) Consider use of a flow diagram		RECORD 13.1: Describe in detail the selection of the persons included in the study (i.e., study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	(Results) Pages 6-18
Descriptive data	14	(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders (b) Indicate the number of participants with missing data for each variable of interest (c) Cohort study - summarise follow-up time (e.g., average and total amount)			(Results) Pages 6-18
Outcome data	15	Cohort study - Report numbers of outcome events or summary measures over time			(Results) Pages 6-18

	Item no.	STROBE items	Location in manuscript where items are reported	STROBE items	Location in manuscript where items are reported
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder -adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorised (c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period			(Results) Pages 6-18
Other analyses	17	Report other analyses done —e.g., analyses of subgroups and interactions, and sensitivity analyses			(Results) Pages 6-18
Discussion Key results		Summarise key results with reference to study objectives			(Discussion) Pages 19-22
Limitations		Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias		RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	(Discussion) Pages 19-21
Interpretation		Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence			(Discussion) Pages 21-22
Generalisability		Discuss the generalisability (external validity) of the study results			(Discussion) Pages 19
Other Information Funding		Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based			Page 22
Accessibility of protocol, raw data and programming code				RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Page 22

