Centre for Community Child Health

murdoch
children's
research
institute

Data Management and Research Workflow Framework

Changing Children's Chances

Nov 2021

The Royal
Children's
Hospital
Melbourne

Data Management and Research Workflow Framework

Version 1.0

Authors

Dr Sarah Gray
Dr Shuaijun Guo
Dr Marnie Downes
Dr Meredith O'Connor
Dr Margarita Moreno-Betancur
Prof Sharon Goldfeld

Changing Children's Chances investigator team

Prof Sharon Goldfeld
Dr Meredith O'Connor
Prof Katrina Williams
Associate Prof Susan Woolfenden
Prof Hannah Badland
Prof Naomi Priest
Dr Margarita Moreno-Betancur
Dr Francisco Azpitarte Raposeiras
Dr Alicia McCoy
Dr Timothy Gilley

Suggested citation: Gray, S., Guo, S., Downes, M., O'Connor, M., Moreno-Betancur, M., Goldfeld, S. and the Changing Children's Chances investigator team. (2021). *Data Management and Research Workflow Framework.* Centre for Community Child Health. Melbourne, Australia.

The Centre for Community Child Health is a research group of the Murdoch Children's Research Institute and a department of The Royal Children's Hospital, Melbourne.

Centre for Community Child Health

The Royal Children's Hospital Melbourne
50 Flemington Road, Parkville
Victoria 3052 Australia
Telephone +61 9345 6150
Email enquiries.ccch@rch.org.au
www.rch.org.au/ccch

Changing Children's Chances is a partnership initiative bringing together leading equity researchers and policy experts from the University of Melbourne, Monash University, The University of New South Wales, Australian National University, The Royal Melbourne Institute of Technology, Loughborough University, Murdoch Children's Research Institute, Beyond Blue, Victorian Health Promotion Foundation, Australian Department of Health, Australian Department of Social Services and Brotherhood of St. Laurence.

Contents

Abbreviations iv

1. Introduction 1

1.1 Ethical Standards 2

1.2 Open Science Practice 2

2. Data sources and access 5

2.1 Longitudinal Study of Australian Children (LSAC) 6

2.2 Multi-Agency Data Integration Project (MADIP) 7

2.3 Other related datasets 8

2.4 Data access 8

2.5 Data security 11

2.6 Data sharing 12

2.7 Data archiving 12

2.8 Retention period 12

3. Planning and paper proposal 13

3.1 Paper proposal 14

3.2 Organising files and documentation 15

4. File creation, data preparation and analysis 19

4.1 Naming conventions 19

4.2 Data preparation 21

4.3 Data analysis 23

4.4 Code review 25

4.5 Derived standard variables 25

5. Paper drafting and reporting results 27

5.1 Manuscript draft 27

5.2 Reporting statistical results 28

6. Paper submission, publication, and outputs 30

6.1 Use the directory template as guidance 30

6.2 Key statements and declarations 30

6.3 Coding shared in the public repository 33

6.4 Research outputs and dissemination 35

7. Summary and conclusions 36

8. References 36

9. Appendices.. 38

9.1 Appendix 1. LSAC data sharing with authorised users through REDCap.................................38

9.2 Appendix 2. Syntax examples of data preparation...39

9.3 Appendix 3. Syntax examples of data analysis 43

9.4 Appendix 4. Syntax examples of exporting results from Stata to Word/Excel 50

Abbreviations

ABS	Australian Bureau of Statistics
ACIR	Australian Childhood Immunisation Records
AEDC	Australian Early Development Census
AIFS	Australian Institute of Family Studies
ARC	Australian Research Council
ATO	Australian Taxation Office
ATP	Australian Temperament Project
CCC	Changing Children’s Chances
CCMS	Child Care Management System
CEBU	Clinical Epidemiology and Biostatistics Unit
DEX	Data Exchange – Family and Community Program
DOMINO	Data Over Multiple Individual Occurrences
DSS	Department of Social Services
E4Kids	Effective Early Educational Experiences
ECEC	Early Childhood Education and Care
FFY	First Five Years
HREC	Human Research Ethics Committee
ICPSR	Inter-university Consortium for Political and Social Research
ITR	Individual Tax Return
LSAC	Longitudinal Study of Australian Children
MADIP	Multi-Agency Data Integration Project
MBS	Medicare Benefits Schedule
MCRI	Murdoch Children’s Research Institute
NAPLAN	National Assessment Program – Literacy and Numeracy
NBE	Neighbourhood Built Environment
NCLD	National Centre for Longitudinal Data
NHMRC	National Health and Medical Research Council
NHS	National Health Survey
NQS	National Quality Standard
PBS	Pharmaceutical Benefits Scheme
PIT	Personal Income Tax
RCH	Royal Children’s Hospital
RD	Registries of Deaths
REDCap	Research Electronic Data Capture
RMIT	Royal Melbourne Institute of Technology

1. Introduction

The purpose of this document is to establish a data management framework for the Changing Children's Chances (CCC) project (2021-2024) "Child health and developmental inequities: Evidence for precision policy." The document aims to promote:

- A standardised approach to data management across team members;
- Open and transparent practices that allow other researchers to follow and replicate the work undertaken;
- Awareness of responsibilities and requirements regarding data security and safety, with easy access to relevant information for CCC project staff; and
- Sharing of data management approaches between researchers, capacity building and collective knowledge for the team and the wider Centre for Community Child Health.

The first phase of the Changing Children's Chances project (2016-2020; hereafter refers to CCC Phase 1) brought together leading national and international child equity researchers to describe children's experiences of disadvantage and its long-lasting impact on their development. In the next phase of the Changing Children's Chances project (2021-2024; hereafter refers to CCC Phase 2), we aim to apply robust epidemiological methodologies to interrogate nationally-representative **existing observational data** to generate evidence that can inform precision policy responses to reduce child health and developmental inequities.

Existing data provide a powerful resource to efficiently, rapidly and robustly generate policy-relevant evidence. We aim to maintain the highest standards of ethical conduct and ensure that data security and participant confidentiality are protected at all times. We are also motivated to conduct research that is informed by the principles of Open Science (an array of practices that promote openness, integrity, and reproducibility in research,[1] see Section 1.2 for further details). By underpinning our research with these values, we hope to increase the transparency and robustness of our data analysis.

In what follows, we provide a guide through each step of the CCC research workflow and consider opportunities to improve openness and reproducibility throughout this workflow, whilst maintaining data security and confidentiality. In practice, this workflow is not always linear, but rather iterative with tasks proceeding in parallel. Of note, this document focuses primarily on **Stata** but corresponding commands exist for **R** users. Hyperlinks to file locations on the MCRI shared drive in this document are only accessible to project staff.

We anticipate that best-practice standards in this space will continue to develop as the project unfolds. Reflecting this, we are approaching this as a working document that can be updated throughout the course of the CCC project.[2]

For further information regarding this document, please contact our project manager Dr Sarah Gray (sarah.gray@mcri.edu.au).

1.1 Ethical Standards

The Murdoch Children's Research Institute (MCRI) is committed to ensuring that all staff behave in a way that promotes public confidence and trust in the organisation. The MCRI Code of Conduct sets out the principles and work values expected for all staff. CCC staff have the responsibility to familiarise themselves with the Code of Conduct and undertake their duties in a manner that is consistent with its policies and procedure.

Ethics approval for the CCC project has been provided by the Royal Children's Hospital (RCH) Human Research Ethics Committee (HREC) (Project Title: Changing Children's Chances: Exploring socio-ecological influences on inequities in children's development; RCH HREC Reference Number: 2019.170). There are no anticipated risks involved to participants for analyses of these existing data.

1.2 Open Science Practice

Open science practice provides an overarching framework for considering how the CCC project can generate science that is open, transparent, and reproducible to enhance and accelerate scientific progress and discovery.[3] We aim to use open science to increase the integrity of our data management procedures by making our decision-making open and traceable, and ensuring that the final research output is replicable when sharing with others. We integrate open science thinking as relevant to each step of the CCC workflow (Figure 1).

Open science refers to the sharing of resources, ideas and places with emphasis on making these publicly and freely available for future use, through three main practices:

- **Preregistration:** A preregistered design includes details of the research design, research questions, and analytic approach. Preregistration does not mean that we cannot deviate from the plan, change course and adapt to new information. Rather, it means there is a record of how and why the plan changed.
- **Open data:** When appropriate, making data publicly available on an open-access repository with corresponding documentation. There are many circumstances, however, in which this is not possible or advisable. For example, sharing data could violate participant confidentiality. In this case, the reason why data has not been shared should instead be disclosed in a manuscript's "data availability statement".
- **Open materials:** Strong documentation of what data were collected and how, and how that data was analysed to create a fully traceable path from data to publication. That is, making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis.

There is a growing interest in the global open science movement where funding bodies, international organisations, governments and institutions have implemented open access policies or guidelines. For instance, the National Health and Medical Research Council (NHMRC) and the Australian Research Council (ARC) have clear open access policies that are consistent with the Australian Government's commitment to open access, open data and intellectual property management.

As an emerging area, there are still challenges to the implementation of open science practice. Open science practice often requires more time and efforts for archiving, documenting, quality controlling of code and data.[4] Open science is still developing and is not yet mainstream across researchers and journals.[4] In the context of analysis on pre-existing data, standards, exemplars of best practice and infrastructure are less developed than for other research methods like clinical trials.

Nevertheless, there are many ways in which open science practices can already be applied to the analysis of observational data within the CCC research program. Table 1 summarises ways that open science thinking can be operationalised at each step of the CCC workflow.

Table 1. Applications of open science practice into CCC research workflow

CCC workflow step	Examples of how open science can be operationalised
Data sources and access	• Ensure only authorised data users have access to the data. • Ensure that data are saved on a password-protected folder. • Be open and transparent about data security,sharing policy and retention period.
Planning and paper proposal	• Use the analysis plan template from the Clinical Epidemiology and Biostatistics Unit (CEBU) to plan and refine data analytic approach before undertaking analysis. • Set up a directory template to organise and manage data and relevant materials. • Make documentation of materials and data explicit and easy-to-find.
File creation, data preparation and analysis	• Use consistent naming conventions for all materials, do/script files and documents. • Have a fully traceable path from the general release data to the paper working dataset. • Document all variables of interest in a spreadsheet, including variable name, label description, informant, and response options. Include decisions about cut-offs and relevant references. • Write annotated do/script files for each step of data analyses including dataset creation, variable creation, multiple imputation and data analysis. Ensure all do/script files are workable, annotated, clear, and can be followed by another researcher. All do/script files enable the replication of each step of data analyses from the source dataset to the output tables and figures. • Document major deviations from the analysis plan in data analysis log. • Create new or use previous standard variable coding documents for long-term data archival, analysis and sharing.
Paper drafting and reporting results	• Include a dot-point summary of major changes to papers and analysis with each draft circulated to co-authors. • Include additional materials (e.g., variable description, additional analyses) in supplementary files when required. • Include a data availability statement in all manuscripts. • Report statistical results according to “ATOMIC” recommendations: Accept uncertainty; be thoughtful, open, and modest; and contribute to an institutional change. Importantly, this includes being aware of the limitations that enhance uncertainty, documenting them and nuancing interpretations accordingly. It also means avoiding dichotomous interpretation of results (e.g., there is an effect / there is

CCC workflow step	Examples of how open science can be operationalised
	not …; there is evidence / there is not…). Rather, discussing the extent of an effect or evidence.
Paper submission, publication, and outputs	• Use the directory template to save all submitted materials. • Use the template of Declarations and Statement (e.g., conflicts of interest) when submitting a paper to a specific journal. • Consider sharing syntax materials from the final version of submitted papers through Open Science Framework or Figshare. • Send the submitted pdf version to all co-authors and save the submitted manuscript as a reference in the shared EndNote library. • Upload bibliographic details of published material (e.g., journal article, conference presentation, theses) to FLoSse Research within 30 days of publication when using Longitudinal Study of Australian Children data.

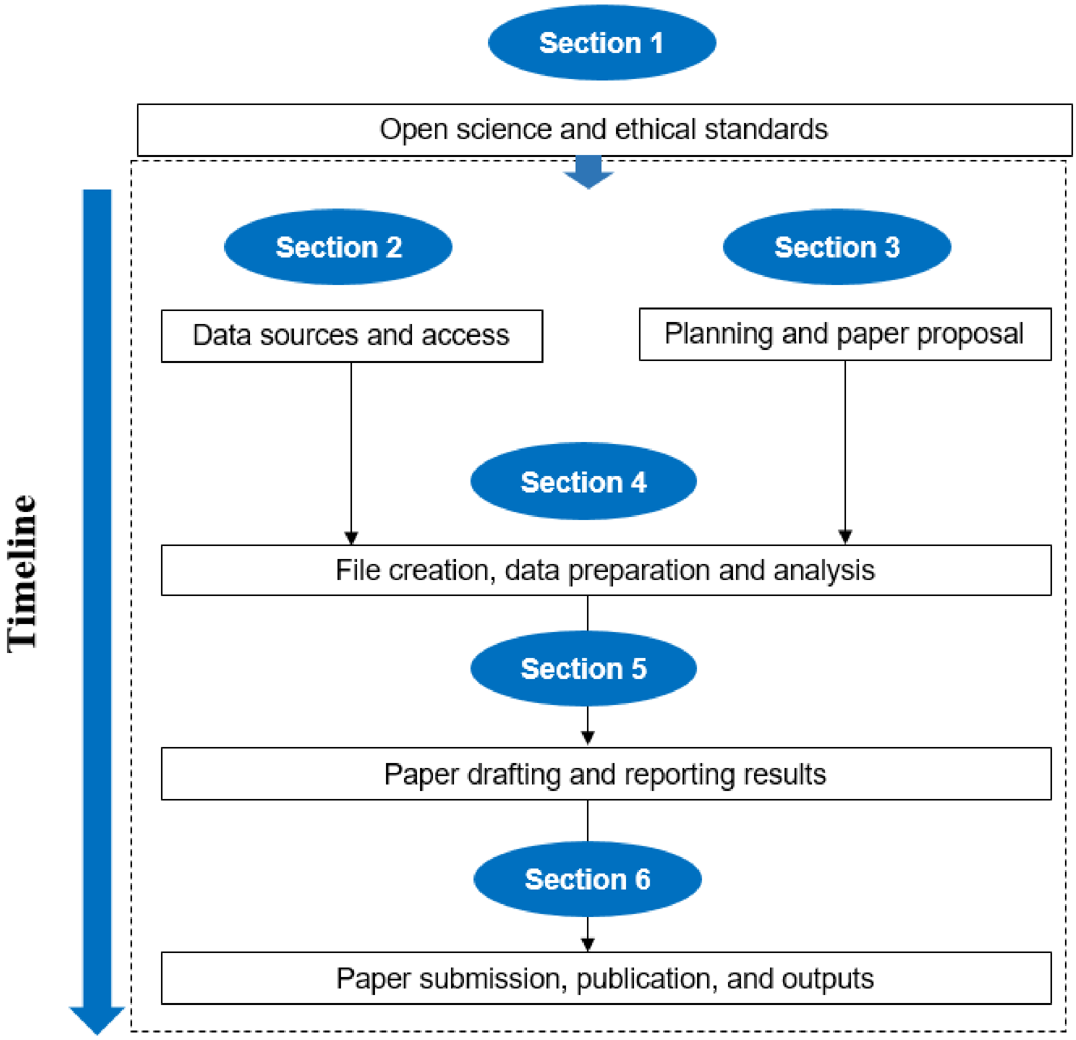

Section 1
Open science and ethical standards
Timeline
Section 2
Data sources and access
Section 3
Planning and paper proposal
Section 4
File creation, data preparation and analysis
Section 5
Paper drafting and reporting results
Section 6
Paper submission, publication, and outputs

Figure 1. The research workflow of Changing Children's Changes Linkage Project

2. Data sources and access

This section provides an overview of two observational datasets that we will use for our project: the Longitudinal Study of Australian Children (LSAC) and the Multi-Agency Data Integration Project (MADIP). Details of these two datasets are presented below. We also summarise details of data access, storage, sharing, archiving, and the retention period of these data.

In brief, LSAC provides great richness of information, while MADIP provides great breadth of coverage of the Australian child population. We will complement LSAC, which provides granular data about children's lives, with MADIP whose population coverage may enhance the generalisability of findings. Together LSAC and MADIP provide comprehensive data on a range of policy levers of interest (e.g., parent's mental health, preschool participation, built environment, and income support) and children's developmental outcomes across multiple domains (i.e., mental health, academic skills, physical health and development), while allowing us to account for the full extent of children's exposure to disadvantage in our investigations. Table 2 shows an overview of key constructs available for analysis across LSAC and MADIP.

Table 2. Overview of key constructs available for analysis across LSAC and MADIP

Domain and indicators	LSAC	MADIP
Exposure to disadvantage		
Sociodemographic	LSAC survey; AEDC	AEDC; CCMS; Census; DOMINO; ITR; MIG; NHS; PIT
Geographic environments	LSAC survey; NBE	AEDC; Census; NHS;
Health conditions	LSAC survey; CheckPoint; MBS; PBS	Census; MBS; NHS; PBS
Risk factors	LSAC survey	Census; NHS
Child outcomes		
Mental health	LSAC survey; AEDC	AEDC; MBS; PBS
Academic/cognitive	LSAC survey; AEDC; NAPLAN	AEDC
Physical health	LSAC survey; CheckPoint; AEDC; MBS; PBS	AEDC; MBS; PBS; RD
Social-level policy levers		
Income support	LSAC survey; Centrelink	DOMINO; NHS
Social housing	LSAC survey	Census; NHS
Housing affordability	LSAC survey; NBE	Census; DOMINO
Rental stress	LSAC survey; NBE	Census; DOMINO
Community-level policy levers		
Built environment	NBE	NHS
School and preschool infrastructure	My school; NBE	NQS
Quality of ECEC	LSAC survey; NBE	NQS
ECEC workforce	-	-
Co-location of health and social services within school or ECEC	-	-
Family-level policy levers		
Parent mental health	LSAC survey; MBS; PBS	MBS; NHS; PBS
Family violence	LSAC survey	-

	Parenting practices	LSAC survey	-
	Home learning environment	LSAC survey	-

AEDC, Australian Early Development Census; CCMS, Child Care Management System; Census, Census of Population and Housing; ECEC, Early Childhood Education and Care; DOMINO, Data Over Multiple Individual Occurrences; ITR, Individual Tax Return; LSAC, Longitudinal Study of Australian Children; MBS, Medicare Benefits Schedule; MIG, Migration data; NAPLAN, National Assessment Program Literacy and Numeracy; NBE, Neighbourhood Built Environment; NHS, National Health Survey; NQS, National Quality Standard; PBS, Pharmaceutical Benefits Scheme; PIT, Personal Income Tax; RD, Registries of Deaths.

A fully searchable data dictionary of LSAC is available here. Members of the research team who are authorised to access MADIP can access the MADIP data dictionary via the Australian Bureau of Statistics (ABS) DataLab.

2.1 Longitudinal Study of Australian Children (LSAC)

LSAC is a nationally representative cohort study that provides comprehensive and longitudinal measures of development from multiple informants and sources including teachers, parents, and the children themselves, as well as direct assessments of the child and linkage to administrative datasets. LSAC comprises two nationally representative prospective cohorts of children - the birth cohort (B-cohort) of 5,107 infants, and the kindergarten cohort (K-cohort) of 4,983 four-year-olds – each of which commenced in May 2004.[5] Data were collected every two years on multiple aspects of child development as well as family and community characteristics. Currently, data from Wave 1 (B-cohort: age 0-1 years; K-cohort: age 4-5 years) to Wave 8 (B-cohort: age 14-15 years; K-cohort: age 18-19 years) are available (Table 3).

Table 3. Age and sample size of children in LSAC cohorts by wave of data collection

Cohorts	Wave 1 (2004)	Wave 2 (2006)	Wave 3 (2008)	Wave 4 (2010)	Wave 5 (2012)	Wave 6 (2014)	CheckPoint (2015)	Wave 7 (2016)	Wave 8 (2018)
B (infant)	0-1 yrs (n=5107)	2-3 yrs (n=4606)	4-5 yrs (n=4386)	6-7 yrs (n=4242)	8-9 yrs (n=4085)	10-11 yrs (n=3764)	11-12 yrs (n=1874)	12-13 yrs (n=3381)	14-15 yrs (n=3127)
K (child)	4-5 yrs (n=4983)	6-7 yrs (n=4464)	8-9 yrs (n=4331)	10-11 yrs (n=4169)	12-13 yrs (n=3956)	14-15 yrs (n=3537)	-	16-17 yrs (n=3089)	18-19 yrs (n=3037)

The capacity of LSAC to address the aims of this research has been enhanced through extensive data linkages. A brief summary of key data linkages within LSAC are as follows:

- **Child Health CheckPoint**: CheckPoint provides state-of-the-art data on the health of Australian children and their parents, including biomarkers (indicators of biological processes) that allow for detailed investigations of how inequity gets under the skin to influence early precursors of adult health problems and noncommunicable diseases (e.g. cardiovascular risk profiles, stress reactivity, and inflammation).[6] Biomarkers are increasingly being used in social sciences to capture information that is not readily available in traditional survey methods.[7] The CheckPoint data will complement the longitudinal physical health measures already available in LSAC (e.g. obesity) in our investigations of physical health outcomes. Data are available for N=1874 children in the LSAC B cohort, collected at 11-12 years of age (between Wave 6 and Wave 7).

- **Neighbourhood Built Environment (NBE)**: Funded by this project, child- and family-relevant neighbourhood built environment indicators will be created and linked to LSAC B cohort Wave 7 (12-13 years of age) at the participant address level, for those living in capital cities and major regional cities. These indicators will provide the basis for fine-grained investigations into the potential of neighbourhood built environment interventions to reduce child inequities. The indicators have been conceptualised and developed through several substantial programs of work at the Centre for Urban Research at RMIT University, being the: NHMRC Centre of Research Excellence in Healthy Liveable Communities, Australian Prevention Partnership Centre National Liveability Study, the Kids in Communities Study (KiCS), and the Australian Early Development Census (AEDC)-NBE pilot study. Permission for linking the neighbourhood built environment data with the LSAC dataset has been granted and will be performed by approved linkage providers, Australian Institute of Family Studies (AIFS).
- **Australian Early Development Census (AEDC)**: AEDC provides information about children's demographic characteristics, preschool experiences, and early developmental outcomes. Teachers complete the AEDC for all Australian students in their first year of compulsory schooling (at about five years of age), using a secure web-based data entry system, across Government, Independent, and Catholic schools.[8] The AEDC is completed every three years, and data are now available from 2009, 2012, 2015, and 2018 with over 250,000 children in each. Linkage was successful for 58.0% (2459/4242) of children at LSAC B cohort Wave 4.
- **National Assessment Program – Literacy and Numeracy (NAPLAN)**: NAPLAN is an Australia-wide direct assessment conducted in schools with all children in Grades 3, 5, 7 and 9,[9] and provides a valuable assessment of multiple dimensions of children's academic progress. Linkage was successful for 89.4% (3651/4085) of children in LSAC B cohort Wave 5 and 99.1% (3351/3381) of children in LSAC B cohort Wave 7.
- **Medicare Australia**: This includes data from the Medicare Benefit Scheme (MBS), the Pharmaceutical Benefit Scheme (PBS) and the Australian Childhood Immunisation Records (ACIR). In Wave 1, 97% of parents of study children gave consent for their children's data to be linked with Medicare Australia data on an ongoing basis. Data from these sources provide details of usage history of MBS, PBS and ACIR services. Linkage was successful for 93% of children in Wave 1.
- **Centrelink welfare**: The Centrelink data are linked with LSAC K cohort Wave 7 and Wave 8, but not for the B cohort.

2.2 Multi-Agency Data Integration Project (MADIP)

CCC is collaborating with the First Five Years: What makes a difference? (FFY) project to access a longitudinal child-centred data asset from the Multi-Agency Data Integration Project (MADIP).[10] MADIP provides large-scale Australian Government administrative data. FFY is led by the Department of Education, Skills and Employment and aims to enhance understanding of the effects of health and socio-economic factors that drive disadvantage with respect to children's early developmental outcomes and identify early childhood policy interventions or protective factors that can improve these outcomes. Specific details of the MADIP dataset are accessible to team members who are authorised to access this data. A brief description of data available in the MADIP can be found via the ABS website, which include:

- AEDC: AEDC includes outcome measures about how well children in their first year of full time school are developing across five important domains. See details in Section 2.1.
- Census of Population and Housing: Census provides key demographics, social and economic data from all people in Australia on Census night, occurring every five years.
- Child Care Management System (CCMS): CCMS contains information on Child Care Benefit for approved child care services. It includes information relating to long day care, after school hours care and before school hours care services.
- Data Exchange (DEX) – Family and Community Program: DEX collects the program performance information that contains de-identified data on clients that receive social services including their demographics and services being delivered.
- Data Over Multiple Individual Occurrences (DOMINO): DOMINO contains information on recipients' demographics, benefits history (e.g., Age Pension and Newstart Allowance), concessions, education (where available) and housing.
- Individual Tax Return (ITR): ITR collects personal tax return information within 16 months of the end of the financial year.
- Medicare Benefits Schedule (MBS): MBS includes information on the usage of Medicare subsidised health care services and corresponding dates.
- Migration: Migration data collects personal information about various migrant types, including permanent, skilled, temporary and other migrant programs, including their demographics and movement over time.
- National Health Survey (NHS): NHS provides data on Australian's health and wellbeing such as medical conditions, health and lifestyle risk factors, mental health and use of health services.
- National Quality Standard (NQS): NQS includes seven quality areas that are important outcomes for children. These data items along with the CCMS use administrative data covering enrolment and attendance of children aged 4-6 years, and their associated carers.
- Personal Income Tax: PIT includes detailed information about taxpayers' occupation and income, employment payments and amounts withheld during a financial year, and all persons with a registered tax file number (TFN) for tax and superannuation purposes.
- Pharmaceutical Benefits Scheme (PBS): PBS includes information about the use of prescription medications, and services subsidised under the PBS and corresponding dates.
- Registries of Deaths (RD): RD hold records for deaths in Australia. The database comprises information about causes of death and other characteristics of the person, such as sex, age at death, area of usual residence and Indigenous status.

2.3 Other related datasets

There are also other additional datasets that are available to team for side studies in our broader data analysis research program, but are not core to CCC Phase 2. Further details can be found in our Phase 1 Data Management Manual.

2.4 Data access

2.4.1 LSAC data access

To become an authorised data user, project staff must sign a copy of the Deed of Confidentiality. Users of the dataset under previous licencing arrangements (organisational or individual) must also complete the National Centre for Longitudinal Data (NCLD) Data Holdings Form, specifying the NCLD datasets they wish to retain or relinquish. Project staff can access blank forms and existing signed copies are here: K:\2. Data\Data management\Data access and reporting\Data access forms\LSAC Wave 8\Application form. Prospective users must also submit a request to access the LSAC data through the ADA Dataverse https://dataverse.ada.edu.au/. Further details about how to apply for LSAC Wave 8 dataset are available to project staff here: K:\2. Data\Data management\Data access and reporting\Data access forms\LSAC Wave 8.

Of note, the current release version of LSAC being used for CCC Phase 2 is 8.0 (Waves 1-8), and there will be updated versions in future. We will keep track of the latest version and download it through the Australian Data Archive. Latest versions of LSAC are saved in our shared drive in two versions: 'Base datasets' and 'Working datasets'. Old versions of LSAC will be kept in the "Archive" subfolder in 'Base datasets'.

LSAC authorised users must immediately notify DSS via email to ada@anu.edu.au in the following situations: change of personal details (e.g. name, phone number or email address, institutional affiliations); change to or addition of research project details; and access to the data is no longer required. Authorised users must also make publicly available all research resulting from the use of the data. Within **30** days of publication or finalisation, authorised users are required to upload bibliographic details of published material to FLoSse Research at flosse.dss.gov.au. FLoSse Research is a publicly available searchable repository of research which uses one or more of DSS longitudinal studies. Types of research that should be uploaded to FLoSse include, but are not limited to: annual reports, journal articles, presentations and conference papers, technical, working papers and reports, theses and student dissertations. More information about this is available in the National Centre for Longitudinal Data Access and Use Guideline.

Currently, authorised LSAC data users (as of Nov 2021) include:

1) Dr Sarah Gray
2) Dr Meredith O'Connor
3) Dr Elodie O'Connor
4) Dr Jun Guo
5) Dr Marnie Downes
6) Dr Margarita Moreno-Betancur
7) Prof Hannah Badland
8) Dr Karen Villanueva
9) Ms Amanda Alderton
10) Ms Rebecca Roberts
11) Ms Fadhillah Norzahari
12) Ms Cindy Pham

2.4.2 LSAC-Neighbourhood Built Environment linkage access

The LSAC-Neighbourhood Built Environment linkage includes child- and family-relevant neighbourhood built environment indicators linked to LSAC B cohort Wave 7 (12-13 years of age) at the participant address level, for those living in Australian capital cities and major regional cities. The linkage is expected to be completed December 2021. Further details to access this dataset will be updated in due course. Currently, researchers who have applied to access this dataset (as of Nov 2021) include:

1) Dr Sarah Gray
2) Dr Meredith O'Connor
3) Dr Elodie O'Connor
4) Dr Jun Guo
5) Dr Marnie Downes
6) Dr Margarita Moreno-Betancur
7) Prof Hannah Badland
8) Dr Karen Villanueva
9) Ms Amanda Alderton
10) Ms Rebecca Roberts
11) Ms Fadhillah Norzahari

2.4.3 MADIP data access

Researchers affiliated with Australian Government or academic research organisations can apply to use MADIP microdata in DataLab for in-depth analysis using a range of statistical software packages. To access the DataLab, researchers need to be approved by the ABS as an Accredited Researcher for MADIP. An accredited researcher must:

- Be able to demonstrate the appropriate knowledge and experience necessary for handling personal information, and demonstrate a commitment to protecting and maintaining the confidentiality of data;
- Be experienced in the use of one of the analytical languages available within the DataLab. The DataLab is a self-service system, and does not include ABS provision of assistance to users with coding or methodological queries for their research;
- Have successfully completed the mandatory DataLab training;
- Have signed an Individual Undertaking and Declaration of Compliance; and
- Work for an institution which has signed a Responsible Officer Undertaking.

Once researchers complete the DataLab training course, they will receive further details and relevant application forms through email communications. Information about the DataLab training can be found here. Current authorised DataLab users are (as of Nov 2021):

1) Dr Sarah Gray
2) Dr Jun Guo
3) Dr Marnie Downes
4) Ms Cindy Pham

2.4.4 AEDC data access

The CCC team has access to the AEDC complete microdata file through an MCRI organisational licence. The Data Manager (Sharon Goldfeld) is responsible and accountable for the dataset's management. The Data Manager manages access to the AEDC data and only permits access to those individuals authorised by the MCRI Organisation. The Data Manager must provide details of this delegation to support@aedc.gov.au once arranged.

Authorised AEDC data users (as of Nov 2021) include:

1) Prof Sharon Goldfeld
2) Dr Sarah Gray
3) Dr Elodie O'Connor
4) Dr Jun Guo
5) Dr Meredith O'Connor
6) Ms Amanda Alderton

2.5 Data security

CCC Phase 2 involves analysis of existing datasets, including data that have already been linked. Participant confidentiality is strictly held in trust by the participating investigators, research staff, and the sponsoring institution and their agents. The study protocol, documentation, data and all other information generated will be held in strict confidence. No information concerning the study or the data will be released to any unauthorised third party, without the prior written approval of the sponsoring institution. Authorised representatives of the sponsoring institution may inspect all documents and records required to be maintained by the investigator.

2.5.1 LSAC data security

We are provided with secure access to existing de-identified LSAC datasets. Data will be stored confidentially (de-identified) in electronic form on the RCH server in a restricted access folder. No name-identified disaggregated information will be used in any publications. Only LSAC authorised data users can access the data. In some cases (e.g., when changing to use a new computer), essential data will be shared through the Research Electronic Data Capture (REDCap) with authorised users (See Section 2.6 Data sharing).

2.5.2 MADIP data security

MADIP data and all relevant analyses will be conducted and saved in the DataLab, which is an interactive data analysis solution available for users to run statistical analyses, using R, SAS, Stata and Python.[11] Controls in the DataLab have been put in place to protect the identification of individuals and organisations. These controls include environmental protections, data de-identification and confidentialisation, access safe guards and output clearance.

DataLab output and working files that are within the DataLab system must not be removed or shared with any other person. This includes not capturing any on-screen information or discussing uncleared DataLab output with users who have not been approved for that microdata or project. To use or share DataLab output outside the DataLab system, authorised users must request clearance of any DataLab output by an ABS officer prior to sharing or disseminating. DataLab output that has been cleared by the ABS may be shared or published and does not need to be securely stored.

2.6 Data sharing

In most cases, LSAC data are shared through the MCRI shared drive with authorised data users (i.e., team members) who work in MCRI. In some cases (e.g., using a new computer), we need to share the LSAC dataset with authorised data users outside MCRI. We will use the REDCap, a web-based software, to transfer our data in a secure and fully transparent way.[12] The REDCap allows for uploading a file up to 128 MB in size. Details of how to share the data securely are described in Appendix 1. MADIP data cannot be shared outside of DataLab.

2.7 Data archiving

According to Jacobs and Humphrey,[13] "data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a project and incorporate a schedule for depositing products over the course of a project's life cycle and the creation and preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method." The CCC project carefully considers archiving at each step of the data lifecycle.

Using the directory template mentioned in Section 3.2, we will archive our datasets and all related documents clearly and safely. Further detailed information on archival preservation can be obtained from the Guide to Social Science Data Preparation and Archiving.[14] In what follows, this document will incorporate and elaborate a plan to address archival considerations at each step.

2.8 Retention period

According to the 2018 Australian Code for Responsible Conduct of Research,[15] all research data will be retained for at least five years from the date of publication. We will delete the LSAC data permanently once the retention period has ended.

3. Planning and paper proposal

In the previous section, we introduced CCC data sources and how to access them and maintain data security. In this section, we move to describe the planning process involved in each proposed CCC paper, specifically focusing on the development of analysis plans.

A good plan for each of our CCC papers keeps our work on track and minimises scope creep. Table 4 presents an overview of planning tasks to be considered at the beginning of a new paper proposal and discussed amongst the CCC project team.

Table 4. An overview of planning tasks when starting a new paper proposal

Planning tasks	How?
Specific goals and publishing plans	1. Begin with the specific research objective: a. What is the research question? b. Why is this important? c. What policy priorities does this link to? 2. Where will we submit the paper?
Scheduling	1. Consider a timeline with target dates for completing key stages of the project (e.g. data collection, cleaning and documenting data, and initial analysis) 2. Set up a reminder for deadlines (e.g. conference abstract, paper submission, external funding) on team members' calendar 3. Note important dates for team members' annual leave
Division of labour	1. Who is responsible for which tasks (e.g. variable extraction, data cleaning, analysis)? 2. If multiple people have access to one document in our shared drive, how do we ensure that only one person is updating the document at a time? 3. Who keeps the documentation up to date? 4. What agreements do team members have about collaboration and joint authorship?
Datasets	1. Which dataset will be used? See details in Section 2.1 and 2.2. 2. Which variables will be used? 3. If it is a multi-cohort study, do we need to apply for access to another dataset? Who will we contact?
Variable names and labels	1. Use consistent conventions for naming and labelling variables, rather than choosing names and labels in an *ad hoc* manner. See details in Section 4.1. 2. When planning variable names, anticipate new variables that could be added later.
Missing data	1. What types of missing data will be encountered, and how will these types be coded? 2. Consideration of why the data are missing (e.g. attrition, refusal, or a skip pattern in the survey) 3. How will we deal with missing data (e.g. deleting incomplete records, multiple imputation)?

Planning tasks	How?
Analysis	1. Complete the CEBU analysis plan template for each proposed paper and consult with CCC investigators to refine. See details in Section 3.1. 2. What types of statistical analyses are anticipated? 3. Who will write the coding and conduct the analysis? 4. What software is needed, and is it locally available? 5. What resources and expertise are available to guide the analysis plan?
Documentation	1. What documentation is needed (e.g. variable description, variable spreadsheet, codebooks)? 2. Who will keep it? In what format? 3. Where will we save all documentation materials?
Backing up, sharing and archiving	1. Who is going to make regular backups of the files (e.g. EndNote library)? 2. Who will we share our coding and logs (e.g. the public, investigators)? How to share? 3. If the research is funded, what requirements does the funding agency have for archiving the data? 4. Long-term preservation should be considered. See details in Section 2.7.
Knowledge translation outputs	1. What types of outputs will we generate? 2. Who are end-users of our research outputs? 3. What resources will we use to enhance our translation capability?

3.1 Paper proposal

A paper proposal is written before the study is conducted and outlines the technical details of a research study.

3.1.1 Why invest time in refining an analysis plan and making this available at least internally?

Research proposals or ‘protocols’ are frequently used in the natural or physical sciences (e.g. clinical trials for new drugs or treatments), and preregistration of these protocols is often a prerequisite for publication. Preregistration is the practice of depositing a research question and study design with a registration service or journal before conducting a scientific investigation.[16] The primary purpose of specifying a research proposal (whether preregistered or not) is to improve the transparency of the findings that are seeking to address a well-defined research question.[17,18]

In recent years, advocates of open science have also promoted the adoption of preregistration of analysis plans within social sciences. Compared with clinical trials, observational studies are particularly subject to publication bias and reporting bias.[19] It is not always easy to distinguish observational studies that are driven by a well-defined pre-specified research question.[20] Preregistration of analysis plans for observational studies has the potential to improve the transparency and rigour of the study design, analysis reporting and interpretation of study findings.[17,18]

3.2.2 How is the analysis plan operationalised in CCC?

As a starting point for each CCC paper, a research proposal will be developed before undertaking analyses. Research proposals will be developed using the CEBU analysis plan template designed for statistical analysis in observational studies,[21] which can be found here: https://doi.org/10.26188/12471380. In brief, this analysis plan will include key information such as background, research questions, and specific analytic approaches. The development of a clear and concise research proposal for each CCC paper will strengthen the quality of our work and increase the efficiency of data analysis and the development of paper manuscripts.

Once developed, a research proposal draft will be circulated to CCC investigators and others as required, for opt-in as co-authors to contribute to the development of the proposal. Based on the feedback from CCC investigators and other co-authors, we will update the proposal with major changes documented as bullet points at the beginning of the document. The final version of the proposal will be circulated to co-authors prior to formal analysis. This process aligns with the open science practice of preregistration and will enhance transparency and accountability by ensuring our analyses for each paper are guided by an agreed-upon plan.

After finalising a research proposal, we will consider sharing the document on a public repository such as Figshare, which will provide a public dated-record of the document. Formal preregistration of the analysis plan could also be considered. A decision on sharing CCC research proposals will be made by co-authors and the project team at the time of drafting a paper. Certain journals such as *Lancet* and *BMJ* are supporting the registration of observational studies and welcoming the inclusion of research protocols.[20,22] Some alternative platforms for preregistration of observational studies are:

- The Open Science Framework Registries;
- ClinicalTrials.gov, which has accommodated the registration of observational studies since its launch in February 2000;[23]
- The World Health Organisation International Clinical Trials Registry Platform;[24]
- UMIN Clinical Trials Registry; and
- Journal publications such as through PLOS Biology and PLOS One, which offer options for peer-review and publication of preregistered research.

3.2 Organising files and documentation

A thoughtful organisation of files makes it easier to document our work, identify files and communicate their location. We will use the following directory template to set up each paper topic (Table 5).

Table 5. An overview plan of a directory template

Project directory	Level 1	Level 2	Level 3	Example files	Purpose
\Paper topic					Paper directory
	\Data files and analyses				Datasets, do-files, and logs
		\Data			Analysis dataset and key variable description
				[example]_[Date].dta	Dataset
				Variable description_[Date].docx	A Word document that briefly describes variables of interest
				Variable codebook_[Date].xlsx	A codebook spreadsheet that includes detailed information of all variables
				\Archive*	Any out-of-date files
		\Syntax			Do-files and log files
				Do file 1_[Date]	Do-files
				Log 1_[Date]	Log files
		\Exported results			Exported results such as tables and figures
				Tables_[Date]	Exported tables
				Graphs_[Date]	Exported figures
	\Manuscript				Peer-reviewed manuscript
		\Drafts			Drafts, modifications and author contribution table
				Draft_ [Date].docx	Drafts of paper
				Modifications_ [Date].docx	Logs of key modifications
				Author contribution table_[Date].docx	Author contribution table
		\Files submitted to [Journal Name]			Files submitted to specific journals
			\Author guideline		Journal author guideline
			\Sample papers		Journal sample papers
			\1st submission		All materials for the 1st submission
			\1st revision		All materials for the 1st revision
	\Meetings				Meeting notes and discussions
				Meeting agenda_[Date].docx	Meeting agenda with the specific date

Project directory	Level 1	Level 2	Level 3	Example files	Purpose
				Meeting minutes_[Date].docx	Meeting minutes with the specific date
	\Published files				Published materials
		\Dataset and coding			Dataset, do-files and logs
		\Figures			Published figures
		\Text			Published files
	\Refs				Bibliography and key references
	\Research snapshot				Research snapshot for communication
		\Drafts			Drafted versions
		\Posted			Published versions

*The folder "Archive" will be set up in all levels for each folder.

Clear documentation is essential through each step of the CCC research workflow to ensure that our work is easily accessible and can be reproduced by others (within the team, or externally). When we are close to submitting a peer-reviewed manuscript, it is good practice to review our documentation, check that we still have the files we used, confirm that all do-files still run, double-check that the numbers in our paper correspond to those in our output, and finally make sure that all this is documented in our research log. The ultimate criterion for whether something should be documented is whether it is necessary for replicating our findings. Table 6 presents an overview of documentation materials for a new CCC paper.

Table 6. An overview of documentation materials for a new CCC paper

Materials	Purpose and how	Details in this document
Annotated do/ script files	1. All do/script files should include the author's name, the name of the document file, and the date it was created. 2. All do/script files should include detailed comments that are echoed in the Stata/R log file and clarify what the output means, where it came from, and how it should be interpreted. 3. If we dichotomise a scale, what was our justification? Mention it in the do/script file.	Section 4.1.2 and Appendix 2, Appendix 3
Research logs	1. Each research log corresponding to each do/script file is the cornerstone of our documentation. 2. It should include dates when work was completed, who did the work, what files were used, and where the materials were located.	Section 4.1.2
Codebook spreadsheets	1. A codebook spreadsheet summarises detailed information on the variables in our dataset. 2. This codebook reflects the final decisions made in collecting and constructing variables.	Section 4.2
Variable description	1. A word document that briefly introduces all variables of interest used in a CCC new paper. 2. This document provides co-authors and other researchers interested in deriving similar variables with an overview of the variable description, including label, response options, and example items.	Section 4.2
Modifications of proposal	1. Major changes to the paper will be sent along with paper iteration to co-authors. 2. Major deviations to the analysis plan will be noted in a research log. 3. Other relevant information such as minor modifications or ideas for future research will be recorded in a separate Word document.	Section 5.1

4. File creation, data preparation and analysis

This section details the information of naming conventions, procedures of creating analysis datasets, generating new variables, as well as standard procedures of data analysis.

4.1 Naming conventions

4.1.1 Documentation files naming conventions

To keep clear records, we will label all necessary documentation files with versions and date in day-month-year format. We will use the following convention for naming our project documents. This convention format includes version numbering and/or initials and date in day-month-year format. For example:

- Data management manual [CCC Phase 2]_v1_15062020.docx

We will also label all document files with initials as additional notes, if necessary, to record the person who provides feedback on that document. For example:

- Data management manual [CCC Phase 2]_v1 _15062020_SG.docx

4.1.2 Do-files and logs naming conventions

We will set up do-files and research logs according to the following four standard procedures of coding. Examples of naming do-files and logs are presented in Table 7. We will include the date (day-month-year format) to record for data analysis history. Any outdated do-files and logs will be saved in the folder of "Archive" in the same folder of these do-files or logs. Examples of syntax are available in Appendix 2.

Table 7. Naming conventions for do-files and logs

Standard procedures of coding	Do-files naming example	Logs naming example
Step 1 - Creation of an analysis dataset: We will create an analysis dataset based on variables of interest.	Do file 1. CREATE DATASET_15062020	Log 1. CREATE DATASET_15062020
Step 2 - Data cleaning and variable creation: We will clean the data and create new variables for subsequent analysis.	Do file 2. DATA CLEANING AND VARIABLE CREATION_15062020	Log 2. DATA CLEANING AND VARIABLE CREATION_15062020
Step 3 - Imputation model: Depending on the percentage of missing values, we will make a decision about whether we will conduct multiple imputation or use the complete data for analysis. If we decide to impute datasets, then we have this coding step for our imputation model. If not, we will skip to Step 4.	Do file 3. IMPUTATION MODEL_15062020	Log 3. IMPUTATION MODEL_15062020

Standard procedures of coding	Do-files naming example	Logs naming example
Step 4 - Analysis: We will conduct analyses using the imputed or complete datasets.	Do file 4. ANALYSIS_15062020	Log 4. ANALYSIS_15062020

4.1.3 Variables and labels naming conventions

There are three basic systems for naming variables: sequential naming (e.g., v1, v2, v3), source naming (e.g., q1, q2a, q2b, q3), and mnemonic naming which uses abbreviations that convey content (e.g., id, female, edu). When creating new variables, we will use mnemonic names as our naming conventions because they partially document the command and the output. In some cases, we will also include a time indicator (e.g., w1sep, w2sep, w7sep) when creating similar variables at different time points in the cohort. The most basic principle for naming variables is that "never change a variable unless you give it a new name." As shown in Table 8, there are other principles to consider when naming a variable. Examples of syntax are available in Appendix 2.

Table 8. Principles for selecting variable names

Principles	Explanation and examples
Anticipate looking for variables	Before you decide on variable names and labels, think about how you will find variables during the analysis. You can use “lookfor [string]”, a Stata command, to search and list all variable names or labels that include the “string” you input.
Use shorter names	Stata only allows names of up to 32 characters but often truncates long names when listing results.
Use clear and consistent abbreviations	Plan your abbreviations and get feedback from a colleague before you finalise them. Then use those abbreviations consistently and keep the list of abbreviations as part of the project documentation. For example, you might use “edu” as an abbreviation for “education”.
Use names that convey content	Names that convey content are easier to use than those that do not. For binary variables, one suggested way is to use a name that indicates the category that is coded as 1. For example, if 0 if male and 1 is female, you could name the variable “female,” not “gender.” (when you see a regression coefficient for gender, is it the effect of being male or being female?)
Be careful with capitalisation	Stata distinguishes between names with the same letters but different capitalisation. For example, “educ”, “Educ”, and “EDUC” are three different variables.
Try names before you decide	Selecting effective names and labels is an iterative process. Continue revising and trying names until you are satisfied.
Keep records for studied variables	It is recommended to back up all studied variables in two versions. One version is a spreadsheet that includes all original variables selected from the data dictionary. The other version is a word document that briefly describes the derived new variables, which you will use for subsequent analysis. See Table 5 for details of where to save.

In addition, every variable should have a variable label. Table 9 describes the main principles for naming variable labels.

Table 9. Principles for naming variable labels

Principles	Explanation and examples
Beware of truncation	A variable label should be long enough to provide the essential information, but short enough that the content can be grasped quickly. Therefore, put the most important information in the first 30 characters of a variable label.
Test labels before you post the file	After creating a set of labels, you always need to check how they work with commands such as “codebook”, “compact” and “tabulate.”
Add notes to variables	This step will help you to easily find how and why you create the variable in the syntax and research log. See Appendix 2 for details.

Value labels assign text labels to the numeric values of a variable. The common rule for value labels is “categorical variables should have value labels unless the variable has an inherent metric.” Two steps are usually needed to create a value label. The first step is to define value label and the second is to assign a defined label to one or more variables. To remove an assigned value label, use "*label values*" without specifying the label. Further details and examples are available in Appendix 2. The key principles for constructing value labels are in Table 10.

Table 10. Principles for constructing value labels

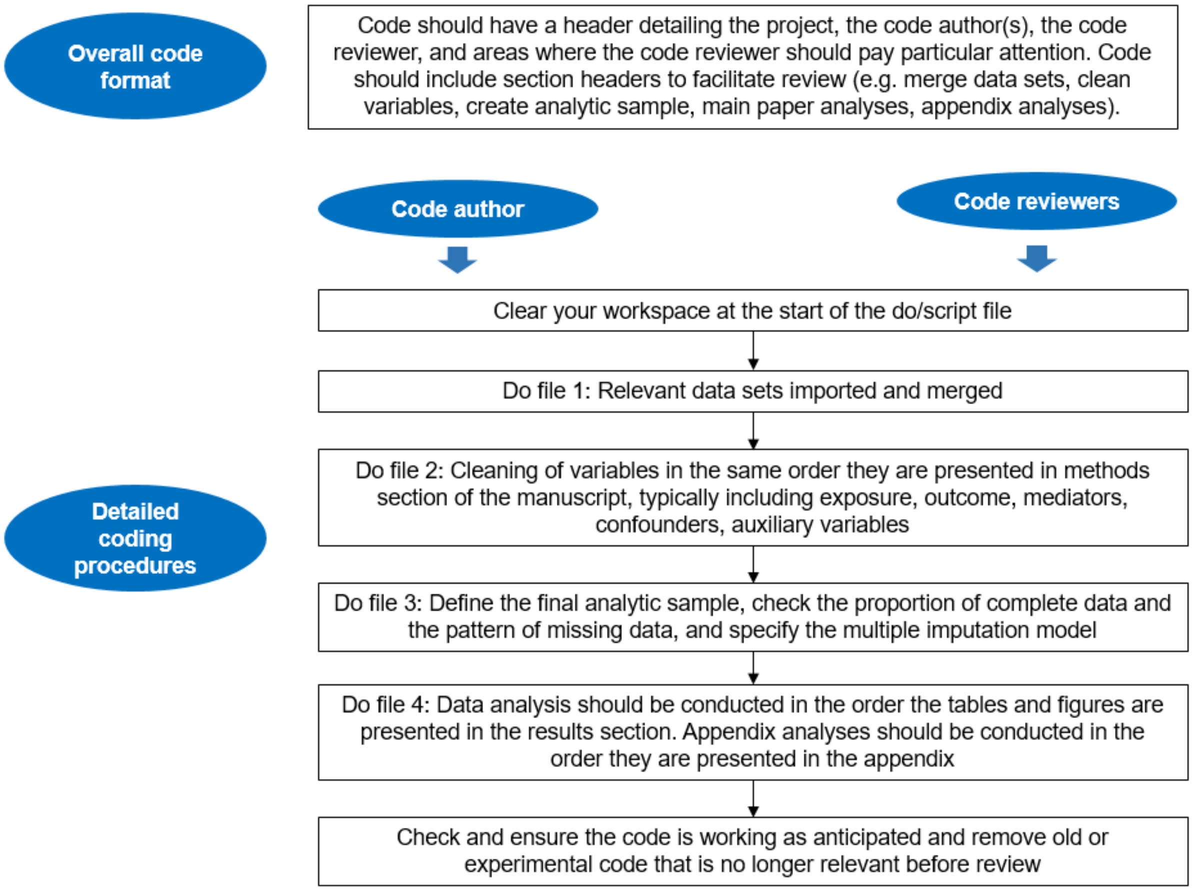
Principles	Explanation and examples
Keep labels short	Value labels should be eight or fewer characters in length.
Include the category number	One way to include numeric values in value labels is to add them when you define the labels. For example “label define defnot 1 1Definite 2 2Probably 3 3ProbNot 4 4DefNot”.
Avoid special characters	Adding spaces and characters such as “. , : _ %” to labels can cause problems with some Stata commands (e.g. hausman), even though “label define” allows you to use these characters in your labels.
Keep track of where labels are used	If a value label is assigned to only one variable, the label definition is recommended using the same name as the variable. If a value label is assigned to multiple variables, the name of the label definition is recommended to begin with a symbol of “L”.

4.2 Data preparation

Once we determine the naming conventions, the next step is to prepare the analysis dataset. We will undertake the following five procedures, as shown in Table 11.

Table 11. Summary of procedures for data preparation

Steps	Details
Step 1 – Identify variables of interest	Once we finalise our research question, we need to identify all study variables from the relevant data dictionary. For example, when searching variables from the LSAC data dictionary, we can use the "Find & Select" function to search keywords and select our samples at a particular time point for either B-cohort or K-cohort.
Step 2 – Build a codebook spreadsheet	After we identify all study variables to be used in a research paper, we will summarise all variables with detailed information in a codebook spreadsheet. This codebook is necessary when cleaning data and creating new variables.
Step 3 – Create do-files, logs and analysis dataset	Once we have the codebook spreadsheet, we will write do-files and create an analysis dataset that includes all new variables derived. We will name each do-file and research log according to name conventions mentioned in Section 4.1.2. Figure 2 details the format and each step of data coding. Main procedures of data coding include: • Create annotated do-files with coding date and analyst name in case of future follow-ups; • Use consistent naming conventions to generate new variables that are easy to read; • Add detailed and traceable notes in do-files such as justifications or references of cut-off values; and • Check each derived variable (e.g. range, observation number) and see whether it is well named and properly labelled.
Step 4 – Set up a table, including all new variables in a Word document	We will generate a table that briefly introduces each new variable used in the analysis dataset, including label description, response options and example items. This table is particularly useful when multiple data users are doing the analysis. In some cases, we can also transform this table as an appendix when submitting a journal paper.
Step 5 – Document all datasets and materials	We will save all materials and datasets in the password-protected folder, using the directory template described in Table 5. We will have a fully traceable path from the general release data to the paper working dataset (i.e., analysis dataset). In some cases, major modifications will be made after circulating the analysis plan with co-authors. We will document these major deviations to the analysis plan in a research log.



Overall code format

Code should have a header detailing the project, the code author(s), the code reviewer, and areas where the code reviewer should pay particular attention. Code should include section headers to facilitate review (e.g. merge data sets, clean variables, create analytic sample, main paper analyses, appendix analyses).

Code author

Code reviewers

Detailed coding procedures

Clear your workspace at the start of the do/script file

Do file 1: Relevant data sets imported and merged

Do file 2: Cleaning of variables in the same order they are presented in methods section of the manuscript, typically including exposure, outcome, mediators, confounders, auxiliary variables

Do file 3: Define the final analytic sample, check the proportion of complete data and the pattern of missing data, and specify the multiple imputation model

Do file 4: Data analysis should be conducted in the order the tables and figures are presented in the results section. Appendix analyses should be conducted in the order they are presented in the appendix

Check and ensure the code is working as anticipated and remove old or experimental code that is no longer relevant before review

Figure 2. The overall and detailed coding format and procedures for code authors and code reviewers (sourced from Vale et al.[25])

4.3 Data analysis

Once we have prepared the analytic dataset, we will conduct analyses using the analysis plan as a guide (see Section 3.1). Generally, we will use the original dataset for descriptive analysis and then use the multiple imputation method to generate imputed datasets. In most cases, we need to consider sampling weights, clustering effects, how to deal with missing data, and sensitivity analysis. We have summarised these commonly used Stata commands in Table 12. Examples of syntax are available in Appendix 3.

Table 12. Summary of commonly used Stata commands when analysing datasets

Command	Purpose
sumtable	creates summary tables by group; this may be treatment groups in a clinical trial or cohort groups in an observational study. The type of summary required for each variable will depend on the data type.

Command	Purpose
svyset	manages the survey analysis settings of a dataset. You use svyset to designate variables that contain information about the survey design, such as the sampling units and weights.
svy	is the survey prefix command that defines the estimation command to be executed.
cluster	executes command(s) on cluster analysis of data.
mdesc	displays the number and proportion of missing values for each variable in varlist.
mvpatterns	lists the missing value patterns of the variables and their frequency.
mi misstable	makes tables to help in understanding the pattern of missing values in your data
mi impute chained	fills in missing values in multiple variables iteratively by using chained equations, a sequence of univariate imputation methods with fully conditional specification of prediction equations.
misum	requires the data to be flong style and calls summarise for each imputed dataset.
mi estimate	runs estimation_command on the imputed mi data, and adjusts coefficients and standard errors for the variability between imputations according to the combination rules by Rubin.
paramed	module to perform causal mediation analysis using parametric regression models.
medeff	is the workhorse command for causal mediation analysis with a variety of data types. For a continuous mediator variable and a continuous outcome variable, the results will be identical to the usual Baron and Kenny method. The command can, however, accommodate other data types, including binary outcomes and mediators.
evalue	performs sensitivity analyses for unmeasured confounding in observational studies using the methodology proposed by VanderWeele and Ding (2017). evalue reports E-values, defined as the minimum strength of association on the RR scale that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates.

After data analysis, we need to move the results from Stata output into our papers or presentations. We can automate much of this work (e.g. exporting tables and high-resolution figures) to reduce errors by manual exporting, and to have a fully traceable path as to how those tables/figures are created. Table 13 summarises some commonly used Stata commands when exporting and saving results.

Table 13. Summary of commonly used Stata commands when exporting and presenting results

Command	Purpose
baselinetable	produces one- and two-way tables of summary statistics for a list of numeric variables.
summtab	computes summary statistics overall and/or across levels of a categorical variable (i.e., the results are stratified by this variable), and compiles them into a nicely formatted, publication-quality table.
putdocx begin	creates document for export
putdocx paragraph	adds paragraph to document
putdocx text	adds text to paragraph
putdocx image	adds image to paragraph
putdocx table	adds table to document
putdocx pagebreak	adds page break to document
putdocx save	closes and saves document

4.4 Code review

Code review is a straightforward technique that can reduce the likelihood of coding bugs. Code review entails a thorough examination of the data cleaning and analysis methods by a team member who was not involved in the initial coding.[25] Typically, the code should be sent for review when the methods and results sections of a paper are nearly finalised. An appropriate code reviewer should be familiar with the software packages, the dataset(s), the methodological approach, interest in the research question and potentially co-authoring the manuscript. The CCC project may adopt a code walkthrough approach to improve our work's reproducibility. The code author will walk the another team member through the code, explaining what is happening in each step (Figure 2). This one-off code review will occur close to the completion of data analyses.

4.5 Derived standard variables

In CCC Phase 1, we generated a series of standard variable documents (e.g. preschool attendance, home reading, mental health service use) to guide how we measure a key construct consistently. These documents are derived when there are multiple variables used to measure a construct, or different options for measuring a constructs, and when the construct is key to CCC analyses (e.g. a mediator of interest, exposure, confounder, or outcome). The standard variables documents summarise indicators relevant to the measurement of a construct, provide a justification for measurement decisions, and syntax for deriving the final agreed-upon variable. Table 14 lists all standard variables derived from our CCC Phase 1 project papers. We will continue updating these working documents when necessary and generate new standard variable documents if required.

Table 14. A brief summary of standard variables derived from CCC Phase 1 Project

Standard variables	Datasets	Age range	Details*
Preschool attendance	LSAC birth cohort	4-5 years (Wave 3)	See document: K:\2. Data\Data management\CCC data management\Standard

Standard variables	Datasets	Age range	Details*
			variables\Preschool\LSAC\Preschool measurement in LSAC_31012019
Home reading	LSAC birth cohort	4-5 years (Wave 3)	See document: K:\2. Data\Data management\CCC data management\Standard variables\Home learning environment\Home learning environment in LSAC_09032018
Child mental health service use	LSAC birth cohort, MBS and PBS	4-7 years (Wave 3 and 4)	See document: K:\2. Data\Data management\CCC data management\Standard variables\Child mental health service use\LSAC plus MBS & PBS\Child mental health service use measurement in LSAC plus MBS & PBS_10022020
Child disadvantage trajectory	LSAC birth cohort	0-1 to 10-11 years (Wave 1 to 6)	See document: K:\2. Data\Data management\CCC data management\Standard variables\Child disadvantage trajectory\Child disadvantage trajectory in LSAC_30072019
Child development outcomes	LSAC birth cohort	10-11 years (Wave 6)	See document: K:\2. Data\Data management\CCC data management\Standard variables\Child development outcomes\Child development outcomes measurement in LSAC_24052019
Child mental health problems	LSAC birth cohort	10-11 years (Wave 6)	See document: K:\2. Data\Data management\CCC data management\Standard variables\Child mental health measurement\Child mental health measurement in LSAC
Child mental health competence and difficulties	AEDC	At 5 years	See document: K:\2. Data\Data management\CCC data management\Standard variables\Mental health competence and difficulties\Mental health competence and difficulties in AEDC
Sociodemographic characteristics	LSAC birth and kindergarten cohorts	0-1 to 10-11 years (Wave 1 to 6)	See document: K:\2. Data\Data management\CCC data management\Standard variables\Sociodemographic characteristics\demographic characteristics in LSAC

*Hyperlinks to K:/ files only accessible to project staff

5. Paper drafting and reporting results

The practices outlined in this section aim to make decisions transparent and traceable throughout the manuscript drafting process.

5.1 Manuscript draft

Paper drafts will be circulated to co-authors who opted-in at the paper proposal stage. Typically, two or three drafts will be circulated to co-authors for feedback at times agreed upon by CI Goldfeld and the project team. We will use the CCC naming conventions to save all versions of paper drafts, with the initials of investigators to indicate versions containing their feedback, to ensure clear tracking of drafts and version control (See Section 4.1).

It is likely that our analysis plan will change and evolve throughout the drafting process, and the final reported analysis will inevitably deviate from the original plan. Therefore, it is important that we keep track of the major analysis decisions and the rationale for those deviations from the original analysis plan. This can also be an extremely helpful when it comes to addressing reviewer comments as knowledge of why a particular approach was used is not lost. Table 15 summarises potential steps that we can consider taking throughout the manuscript drafting process to ensure our analysis decisions are transparent and justified.

Table 15. Key principles of the manuscript drafting process

Principle	Explanation
Principle 1: Keep the original research proposal traceable.	Save the original research proposal in a clear location on our shared drive (or a public repository) so that the planned research is traceable.
Principle 2: Report major results of all pre-specified work.	Analyses outlined in the research proposal will sometimes not be reported in the final paper. We will keep records of results of major pre-planned analyses in an analysis log. We will consider reporting these results in supplementary files if they are relevant to the interpretation of the reported findings and if their exclusion reduces the robustness and transparency of the work.
Principle 3: Clearly label any unplanned analyses.	Unplanned analyses commonly occur. Diverging from our planned analysis does not invalidate our analysis plan. If changes from the original plan are made, these will be noted in the analysis log.
Principle 4: Include a “Transparent Changes” document for any deviations that occurred from the original paper proposal.	Save all major changes from the original proposal (e.g. methodological or analytic changes, or changes in aims and scope) and justifications for these changes in a separate Word document, according to the directory template. Send a dot-point summary of major changes to co-authors with each draft.

5.2 Reporting statistical results

The p-value is the most commonly encountered inferential statistic and one of the most frequently misunderstood and misinterpreted statistics in the literature.[26,27] In 2016, the American Statistical Association[28] released a "Statement on Statistical Significance and p-values" with six principles underlying the proper use and interpretation of the p-value: (1) P-values can indicate how incompatible the data are with a specified statistical model; (2) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone; (3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold; (4) Proper inference requires full reporting and transparency; (5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result; and (6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The CCC project will aim to report statistical results consistent with the above principles. In 2019, Wasserstein et al.[29] summarised five recommendations for reporting statistical results, known as "ATOMIC": **A**ccept uncertainty; be **t**houghtful, **o**pen, and **m**odest; call for an **i**nstitutional **c**hange (see Table 16 for how these may be applied in CCC). For example, throughout the course of CCC Phase 2, we will avoid using the term "statistically significant", "significantly different", "p<0.05" and "non-significant" (See journal author guidelines International Journal of Epidemiology, Epidemiology, American Journal of Epidemiology). More examples of statistical results description without using "statistically significant" are available to project staff here: K:\2. Data\Data management\Analytic resources and examples\Moving to a World Beyond p 0.05.

Table 16. The ATOMIC recommendations to move to a world beyond p<0.05

Recommendation	Explanation and how to apply it into CCC
Accept uncertainty	• We will countenance uncertainty in all statistical conclusions, seeking ways to quantify, visualise and interpret the potential for bias due to design limitations and imprecision due to chance. • Reporting and interpreting point and interval estimates will be routine of CCC papers. We will accompany every point estimate with a measure of its uncertainty (e.g. standard error or interval estimate).
Be thoughtful	• Begin with clearly expressed research objectives • Modelling assumptions (e.g., model specification, handling of missing data) will be sufficiently documented. Are modelling assumptions understood? Are these assumptions valid? Do the key results hold up when other choices are made? • When interpreting the statistical results, consider the scientific context, prior evidence and practical importance, e.g. what do we already know? What magnitude of the effect, odds ratios etc. are practically important in the context of the research? If there is a definition of a meaningful effect size, communicate this up-front before data are analysed.

Recommendation	Explanation and how to apply it into CCC
Open	• Be open to open science practices. • Base judgements (e.g., developing models, interpreting results) on evidence and careful reasoning, and seek expert judgement wherever possible to eliminate potential sources of bias. • We will follow the analysis plan and conduct all planned analyses. If major modifications are made, we will keep a traceable record of these changes. • Report p-values as a continuous and descriptive statistic (e.g., say “p = 0.03” instead of “statistically significant” or “p < 0.05”) and interpret p-values in light of context (sample size and meaningful effect size). • Rigorously document decisions (e.g. cut-off points of key variables, statistical modelling, selection of confounders) and use sensitivity analysis for a better understanding of the impact of choice.
Modest	• Take the role of a neutral judge, rather than an advocate for any hypothesis. This can be done by testing alternative hypotheses, discussing practical implications of endpoints of every interval estimate (not only whether it contains the null). • Be careful not to overreach in the generalizability of claims. Be aware of and acknowledge the limitations of methods in the main text and abstract. • Remember that one study is rarely enough – seek replication and provide sufficient information for replication.
Institutional change	• As authors, we can support institutional change by submitting well-designed studies for publication regardless of findings, referring to the ASA statement when submitting a paper or responding to reviewers, and challenge editors and reviewers when they judge our results because of p-values.

6. Paper submission, publication, and outputs

This section summarises practices to ensure efficient submission of papers and documentation of outputs.

6.1 Use the directory template as guidance

Prior to submission, we will double check that the numbers in our paper correspond to those in our Stata output and make sure that all results are documented in the research log. When submitting CCC papers, we will use the directory template specified in Section 3.2 to manage our documents and shared data coding. For example, if we submitted our paper to *JAMA Pediatrics*, all documents will be saved using the following template shown in Table 17. All shared datasets and coding will be saved in the folder "Published files".

Table 17. Example of project directory for paper submission and publication

Project directory	Level 1	Level 2
\Paper topic		
	\Files submitted to *JAMA Pediatrics*	
		\Archive
		\Author guideline
		\Sample papers
		\1st submission
		\1st revision
	\Published files	
		\Archive
		\Dataset and coding
		\Figures
		\Text

6.2 Key statements and declarations

The following statements and declarations are included in submitted papers (Table 18).

Table 18. Necessary statements and declarations

Sections	Example of content
CCC ethics statement	Ethics approval for secondary data analysis has been provided by the Royal Children's Hospital Human Research Ethics Committee (Project Title: Changing Children's Chances: Exploring socio-ecological influences on inequities in children's development; RCH HREC Reference Number: 2019.170; see modification submitted 13/07/2021).
LSAC ethics statement	The research methodology and survey content of Growing Up in Australia is reviewed and approved by the Australian Institute of Family Studies Ethics Committee. Details are available here about the ethics application numbers (where available) and dates for each wave of LSAC.

Sections	Example of content
MADIP ethics statement	To be updated
Conflict of interest	When no conflicts of interest are identified, e.g.: The authors have no conflicts of interest relevant to this article to disclose.
Financial disclosures	When there are no financial disclosures, e.g.: The authors have no financial relationships relevant to this article to disclose.
Funding	For example (to be updated accordingly): This work was supported by the Australian Research Council Linkage Projects [LP190100921] and was supported by the Victorian Government's Operational Infrastructure Support Program. Prof Goldfeld is supported by Australian National Health and Medical Research Council (NHMRC) Practitioner Fellowship 1155290. Dr O'Connor is supported by the Melbourne Children's LifeCourse initiative, funded by a Royal Children's Hospital Foundation Grant (2018-984). Dr Moreno-Betancur is supported by Australian Research Council Discovery Early Career Award DE190101326. Prof Badland is supported by an RMIT University VC Senior Research Fellowship. Prof Priest was supported by a NHMRC Career Development Fellowship APP1123677. Dr Francisco Azpitarte also acknowledges financial support from the Spanish State Research Agency and the European Regional Development Fund (ECO2016-76506-C4-2-R).The Changing Children's Chances investigator team oversees this program of work, and includes Prof Sharon Goldfeld, Dr Meredith O'Connor, Prof Katrina Williams, A/Prof Sue Woolfenden, Prof Hannah Badland, Prof Naomi Priest, Dr Margarita Moreno-Betancur, Dr Francisco Azpitarte Raposeiras, Dr Alicia McCoy, and Dr Timothy Gilley.
Role of the funder	The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.
LSAC acknowledgements	This paper uses unit record data from Growing Up in Australia, the Longitudinal Study of Australian Children (LSAC). LSAC is conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this paper, however, are those of the authors and should not be attributed to the Australian Government DSS or any of DSS' contractors or partners. DOI: 10.26193/F2YRL5
MADIP acknowledgements when using ATO data	The results of these studies are based, in part, on tax data supplied by the ATO to the ABS under the *Taxation Administration Act 1953*, which requires that such data is only used for the purpose of administering the *Census and Statistics Act 1905*. Any discussion of data limitations or weaknesses is in the context of using the data for statistical purposes, and is not related to the ability of the data to support the ATO's core operational requirements.

Sections	Example of content
	Legislative requirements to ensure privacy and secrecy of these data have been followed. For access to MADIP data under Section 16A of the *ABS Act 1975* or enabled by section 15 of the *Census and Statistics (Information Release and Access) Determination 2018*, source data are de-identified and so data about specific individuals has not been viewed in conducting this analysis. In accordance with the *Census and Statistics Act 1905*, results have been treated where necessary to ensure that they are not likely to enable identification of a particular person or organisation.
MADIP acknowledgements when using Home Affairs (migration) data	The results of these studies are based, in part, on migration data supplied by Home Affairs to the ABS under the Australian Border Force Act 2015, which requires that such data is only used for the purposes of the Census and Statistics Act 1905 or performance of functions of the ABS as set out in section 6 of the Australian Bureau of Statistics Act 1975. Any discussion of data limitations or weaknesses is in the context of using the data for statistical purposes, and not related to the ability of the data to support Home Affairs' core operational requirements. Legislative requirements ensure privacy and secrecy of these data are followed. For access to MADIP data under Section 16A of the ABS Act 1975 or enabled by section 15 of the Census and Statistics (Information Release and Access) Determination 2018, source data are de-identified and so data about specific individuals has not been viewed in conducting this analysis. In accordance with the Census and Statistics Act 1905, results have been treated where necessary to ensure that they are not likely to enable identification of a particular person or organisation.
Spatial data and maps statement	For publications using built environment data: "Spatial data have been provided by the Australian Urban Observatory and Healthy Liveable Cities Group, Centre for Urban Research, RMIT University with funding support provided through the Australian Prevention Partnership Centre, NESP Clean Air and Urban Landscapes Hub and NHMRC Centre of Research Excellence in Healthy, Liveable Communities. Any publications utilising the data are not necessarily the view of or endorsed by RMIT University or the Healthy Liveable Cities Group. RMIT excludes all liability for any reliance on the data." For any maps that are included in the publication: "Spatial data have been provided by the Australian Urban Observatory and Healthy Liveable Cities Group, Centre for Urban Research, RMIT University."
Data sharing statement	Data are not publicly accessible; for LSAC data access queries or requests, see: https://growingupinaustralia.gov.au/; for MADIP data

Sections	Example of content
	access queries or requests, see: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/1900.0main+features5Australia#MADIP
Author contributions	For example (to be updated accordingly): Prof Goldfeld obtained funding, conceptualised, and designed the study, and critically reviewed the manuscript for important intellectual content. Dr Meredith O'Connor, Dr Mensah, Dr Gray, and Dr Elodie O'Connor conceptualised and designed the study, drafted the initial manuscript, and critically reviewed the manuscript for important intellectual content. Dr Moreno-Betancur and Dr Guo conceptualised and designed the study, conducted analysis, drafted the initial manuscript, and critically reviewed the manuscript for important intellectual content. A/Prof Woolfenden, Prof Williams, Dr Kvalsvig, Prof Badland, Dr Azpitarte, and Dr Chong conceptualised and designed the study, and critically reviewed the manuscript for important intellectual content. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

6.3 Coding shared in the public repository

Given the data sharing restriction policy, we will not share our analysis dataset in the public repository. However, to increase transparency and scrutiny of analytic procedures, we can consider sharing our syntax materials internally or in a public repository where appropriate through Open Science Framework or Figshare. We have also summarised alternative options available to share coding in the public data repository (see Table 19).[30] Decisions about when and where to share coding will be made at the time of publication, following discussions with CI Goldfeld and the CCC team.

Table 19. Governance Attributes of General Data Repositories

Repository	Data type	Who can access it?
Open Science Framework	General science content, including data, materials, and code	Access may be public (depositors select from common licenses or upload their own) or private (accessible only to the depositor, contributors to the project or component, and users with a view-only link generated by the depositor).
Figshare	Research data and other outputs (figures, theses, etc.) from any science field, in any file format, up to 5 GB.	Data may be marked as private (accessible only to the uploader while logged in or to other people via a privately shared link) or public.

Repository	Data type	Who can access it?
Databrary	Video, audio, and related metadata in the developmental and learning sciences	Five tiers are available: public, authorised users (data are available to users who are registered and have signed an access agreement co-signed by their home institution), excerpts (data are available to authorised users, who may show clips during presentations; see Gilmore, Kennedy, & Adolph, 2018, this issue), private (data are available only to collaborators), and unreleased (data are accessible only by the depositor).
Dryad	Content associated with scholarly research documents that are published, in press, or under review	By default, data and other content associated with a scholarly research document are made public.
Harvard Dataverse	Quantitative and qualitative data in any format, from any discipline	Although metadata are always open access, files themselves may be restricted use, in which case down-loaders must be registered users.
Inter-university Consortium for Political and Social Research (ICPSR)	Social and behavioral research data of all file types	The vast majority of ICPSR data holdings are public-use files with no restrictions on access. However, in some cases, ICPSR provides vetted researchers and sponsor-supervised students access to restricted-use data versions that retain confidential or sensitive data.
OpenfMRI	All forms of neuroimaging data that include Magnetic resonance imaging (MRI) images and associated data	Unless otherwise noted, data are available under the Creative Commons CC0 1.0 license.
openICPSR	Social and behavioral research data of all file types	Self-publishers choose to either make the data available for immediate public download or to restrict access. If access is restricted, users must apply for access and pay an administrative fee.
OpenNeuro	Neuroimaging data in Brain Imaging Data Structure format	Uploaded data are private (i.e., only collaborators can view and edit the data) for a limited time and then become public.
Zenodo	Any research output (including multimedia) from any field; up to 50 GB per dataset	Data may be marked as open, embargoed (data will become public at the end of a specified timeframe), restricted (access is available only with the permission of the depositor), or closed.

6.4 Research outputs and dissemination

After submitting a paper to a specific journal, we will send the submitted pdf version to all co-authors and save the submitted manuscript as a reference in the shared EndNote library. We will also comply with relevant data access policies. LSAC authorised data users are required to make publicly available all research (e.g., journal article, presentations and conference papers, working papers and technical reports) resulting from the use of the data. Within **30 days** of publication or finalisation, LSAC authorised users are required to upload bibliographic details of published material to FLoSse Research, which is a publicly available searchable repository of research which uses one or more of DSS longitudinal studies. MADIP authorised data users are required to cooperate with any ABS audit directions relating to DataLab usage, code and output. Outside of the DataLab environment, an MADIP authorised data user can only share outputs that have been cleared by an ABS officer.

7. Summary and conclusions

In summary, we are aiming to apply open science into the CCC research workflow to increase clarity, reproducibility, and transparency of our research practice, whilst maintaining data security and confidentiality. Open science practice serves as an overarching framework throughout this document to ensure each step of our work is traceable and replicable when sharing with others. Along with the movement from an era of “Publish or Perish” towards an era of “Visible or Vanish”, we are expecting to promote the sharing of best practices between researchers, capacity building and collective knowledge for the team and the wider academic community.

8. References

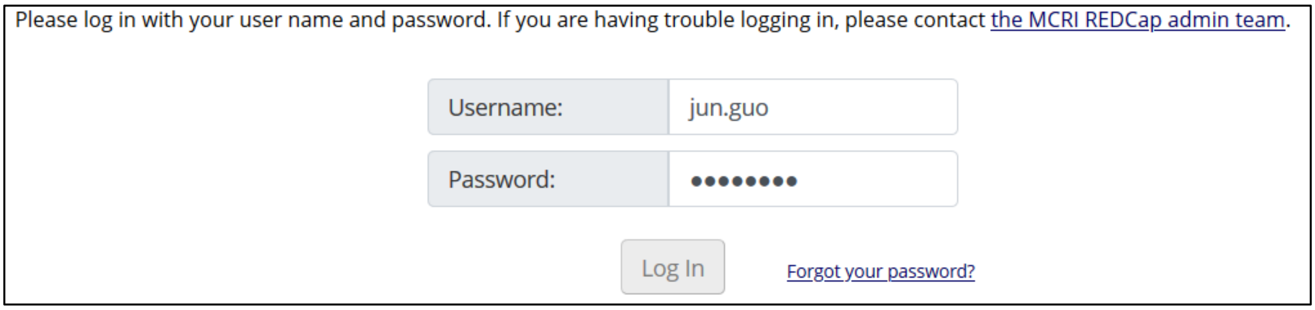
1. Banks GC, Field JG, Oswald FL, et al. Answers to 18 questions about open science practices. *Journal of Business and Psychology.* 2019;34(3):257-270.
2. Long JS. *The workflow of data analysis using Stata.* Stata Press College Station, TX; 2009.
3. Klein O, Hardwicke TE, Aust F, et al. A practical guide for transparency in psychological science. *Collabra: Psychology.* 2018;4(1):1-15.
4. Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. *PLoS biology.* 2019;17(5):e3000246.
5. Soloff C, Lawrence D, Johnstone R. *LSAC Technical paper No. 1. Sample design.* Melbourne, Australia: Australian Institute of Family Studies; 2005.
6. Wake M, Clifford S, York E, et al. Introducing Growing Up in Australia's Child Health CheckPoint: A physical health and biomarkers module for the Longitudinal Study of Australian Children. *Family Matters.* 2014;94:15-23.
7. Shahaeian A, Wang C, Tucker-Drob E, Geiger V, Bus AG, Harrison LJ. Early Shared Reading, Socioeconomic Status, and Children's Cognitive and School Competencies: Six Years of Longitudinal Evidence. *Scientific Studies of Reading.* 2018;22(6):485-502.
8. Brinkman S, Gregory T, Goldfeld S, Lynch J, Hardy M. Data resource profile: The Australian Early Development Index (AEDI). *International Journal of Epidemiology.* 2014;43(4):1089-1096.
9. Australian Curriculum Assessment and Reporting Authority. *NAPLAN achievement in reading, persuasive writing, language conventions and numeracy: National report for 2008.* Sydney, Australia: Australian Curriculum Assessment and Reporting Authority;2008.
10. Australian Bureau of Statistics. Multi-Agency Data Integration Project (MADIP). 2019; https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Statistical+Data+Integration+-+MADIP. Accessed 25 July, 2017.
11. Parker T. The DataLab of the Australian bureau of statistics. *Australian Economic Review.* 2017;50(4):478-483.
12. Obeid JS, McGraw CA, Minor BL, et al. Procurement of shared data instruments for research electronic data capture (REDCap). *Journal of biomedical informatics.* 2013;46(2):259-265.
13. Jacobs JA, Humphrey C. Preserving research data. *Communications of the ACM.* 2004;47(9):27-29.
14. The Inter-university Consortium for Political and Social Research (ICPSR). *Guide to Social Science Data Preparation and Archiving–Best Practice Through the Data Life Cycle.* Michigan, USA: ICPSR;2012.
15. National Health and Medical Research Council. *Australian Code for the Responsible Conduct of Research.* Canberra, Australia: NHMRC;2018.
16. PLOS. Open Science: Preregistration. 2020; https://plos.org/open-science/preregistration/. Accessed June 12, 2020.
17. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proceedings of the National Academy of Sciences.* 2018;115(11):2600-2606.
18. Rushton L. Should protocols for observational research be registered? *Occupational and Environmental Medicine.* 2011;68:84-86.
19. Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. Publication bias in clinical research. *The Lancet.* 1991;337(8746):867-872.
20. Loder E, Groves T, MacAuley D. Registration of observational studies. The next step towards research transparency. *BMJ.* 2010;340:375-376.

21. MARGARITA M-B. *Analysis plan template for life-course cohort studies.* University of Melbourne; 2020.
22. Editors. Should protocols for observational research be registered? *Lancet.* 2010;375:348.
23. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? *Canadian Medical Association Journal.* 2010;182(15):1638-1642.
24. World Health O. International standards for clinical trial registries: the registration of all interventional trials is a scientific, ethical and moral responsibility. 2018.
25. Vable AM, Diehl SF, Glymour MM. Code Review as a Simple Trick to Enhance Reproducibility, Accelerate Learning, and Improve the Quality of Your Team's Research. *American Journal of Epidemiology.* 2021.
26. Cohen J. The earth is round (p<. 05). *American Psychologist.* 1994;49(12):997.
27. Berry D. A p-value to die for. *Journal of the American Statistical Association.* 2017;112(519):895-897.
28. Pernet C. Null hypothesis significance testing: a short tutorial. *F1000Research.* 2016;4:621.
29. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p<0.05". *The American Statistician.* 2019;73(sup1):1-19.
30. Meyer MN. Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science.* 2018;1(1):131-144.
31. Soloff C, Lawrence D, Misson S, Johnstone R. *LSAC Technical paper No. 3. Wave 1 weighting and non-response.* Melbourne, Australia: Australian Institute of Family Studies; 2006.
32. Norton A, Monahan K. *LSAC Technical paper No. 15. Wave 6 weighting and non-response.* Melbourne, Australia: Australian Bureau of Statistics; 2015.
33. Australian Institute of Family Studies. *The Longitudinal Study of Australian Children: An Australian Government Initiative. Data User Guide - October 2020.* Melbourne, Australia: Australian Institute of Family Studies;2020.
34. Clifford S, Davies S, Gillespie A, et al. *Longitudinal Study of Australian Children's Child Health CheckPoint Data User Guide.* Melbourne: Murdoch Children's Research Institute;2020.
35. Williams RL. A note on robust variance estimation for cluster-correlated data. *Biometrics.* 2000;56(2):645-646.
36. Arceneaux K, Nickerson DW. Modeling certainty with clustered data: A comparison of methods. *Political Analysis.* 2009;17(2):177-190.
37. Greenland S. Principles of multilevel modelling. *International Journal of Epidemiology.* 2000;29(1):158-167.
38. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine.* 2011;30(4):377-399.

9. Appendices

9.1 Appendix 1. LSAC data sharing with authorised users through REDCap

Step 1: Log in to REDCap using your personal username and password from this website: https://redcap.mcri.edu.au/.

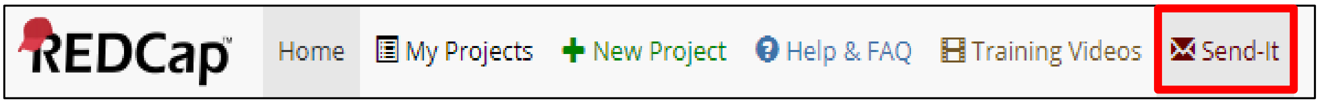


Please log in with your user name and password. If you are having trouble logging in, please contact the MCRI REDCap admin team.

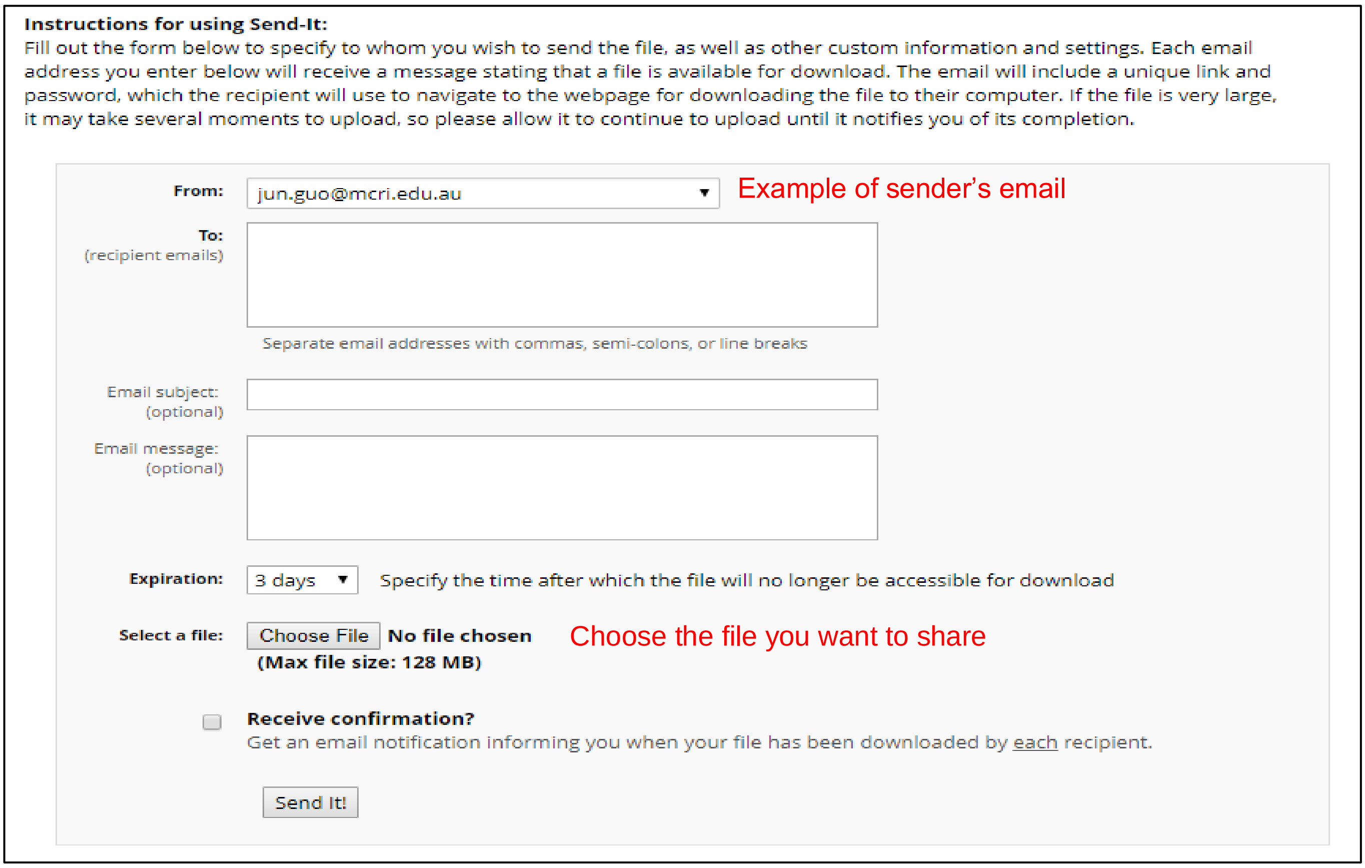
Username: jun.guo

Password: ••••••••

Step 2: Click the button "Send It" in the top menu on the main page.



Step 3: Fill out the form below to specify to whom you wish to share file and then click "Send it" button once the form is completed. The recipient (can be anybody, not only for those with MCRI credentials) will receive two emails which include a weblink and password to download the shared file.



Instructions for using Send-It:
Fill out the form below to specify to whom you wish to send the file, as well as other custom information and settings. Each email address you enter below will receive a message stating that a file is available for download. The email will include a unique link and password, which the recipient will use to navigate to the webpage for downloading the file to their computer. If the file is very large, it may take several moments to upload, so please allow it to continue to upload until it notifies you of its completion.

From: jun.guo@mcri.edu.au ▾ Example of sender's email

To: (recipient emails)

Separate email addresses with commas, semi-colons, or line breaks

Email subject: (optional)

Email message: (optional)

Expiration: 3 days ▾ Specify the time after which the file will no longer be accessible for download

Select a file: Choose File **No file chosen** Choose the file you want to share
(Max file size: 128 MB)

□ **Receive confirmation?**
Get an email notification informing you when your file has been downloaded by each recipient.

9.2 Appendix 2. Syntax examples of data preparation

Naming do-files and research logs

As mentioned in Section 4.1.2, we will create five do-files and corresponding research logs for each research paper. Box 1 shows how we name a do-file and a research log.

Box 1. Example of naming do-files and log files

```
*Create do file 1 and save it as “Do file 1. CREATE DATASET_15062020.do”
*Create log file 1
log using "K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data
files and analyses\Analysis datasets\Working syntax and output\Log 1. CREATE
DATASET_15062020.smcl"

*Create do file 2 and save it as “Do file 2. DATA CLEANING AND VARIABLE CREATION_15062020.do”
*Create log file 2
log using " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data
files and analyses\Analysis datasets \Working syntax and output\Log 2. DATA CLEANING AND
VARIABLE CREATION_15062020.smcl"

*Create do file 3 and save it as “Do file 3. IMPUTATION MODEL_15062020.do”
*Create log file 3
log using " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data
files and analyses\Analysis datasets \Working syntax and output \Log 3. IMPUTATION
MODEL_15062020.smcl"

*Create do file 4 and save it as “Do file 4. ANALYSIS_15062020”
*Create log file 4
log using " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data
files and analyses\Analysis datasets \Working syntax and output \Log 4. ANALYSIS_15062020.smcl"

*Create do file 5 and save it as “Do file 5. RUN ALL DO FILES_15062020.do”
log using " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data
files and analyses\Analysis datasets \Working syntax and output \Log 5. RUN ALL DO
FILES_15062020.smcl"

do " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data files and
analyses\Analysis datasets\Working syntax and output \ Do file 1. CREATE DATASET_15062020.do"

do " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data files and
analyses\Analysis datasets\Working syntax and output \ Do file 2. DATA CLEANING AND VARIABLE
CREATION_15062020.do"

do " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data files and
analyses\Analysis datasets\Working syntax and output \ Do file 3. IMPUTATION
MODEL_15062020.do"

do " K:\1. Studies\Core CCC studies\Mediators of effect of disadvantage on outcomes\Data files and
analyses\Analysis datasets\Working syntax and output \ Do file 4. ANALYSIS_15062020.do"
```

Merging datasets

In most cases, we need to merge data from different waves in the LSAC. Data files can be combined with the *merge* command, which joins corresponding observations from the dataset currently in memory (called the master dataset) with those from *filename.dta* (called the using dataset), matching on one or more key variables.

Box 2. Merging LSAC datasets wave 1 to 7

```
*Set a high maximum number of variables in dataset given merging of large datasets
set maxvar 32767, perm

*Create working dataset using LSAC Wave 1 data (0-1 years)
use "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb0.dta", clear

*Merge in LSAC Wave 2 data (2-3 years)
merge 1:1 hicid using "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb2.dta"
rename _merge mergew2
tab mergew2

*Merge in LSAC Wave 3 data (4-5 years)
merge 1:1 hicid using "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb4.dta"
rename _merge mergew3
tab mergew3

*Merge in LSAC Wave 4 data (6-7 years)
merge 1:1 hicid using "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb6.dta"
rename _merge mergew4
tab mergew4

*Merge in LSAC Wave 5 data (8-9 years)
merge 1:1 hicid using "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb8.dta"
rename _merge mergew5
tab mergew5

*Merge in LSAC Wave 6 data (10-11 years)
merge 1:1 hicid using "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb10.dta"
rename _merge mergew6
tab mergew6

*Merge in LSAC Wave 7 data (12-13 years)
merge 1:1 hicid using "K:\2. Data\Working datasets\LSAC Wave 7 release\lsacgrb12.dta"
rename _merge mergew7
tab mergew7

*Save new dataset
save "[your path here]\Analysis datasets\Working dataset.dta", replace
```

Appending datasets

Sometimes we also need to append another dataset to a working dataset, so that we can add observations to the existing variables. The *append* command appends Stata-format datasets stored on disk to the end of the dataset in memory. If a variable is a string in one dataset and numeric in the other, Stata issues an error message unless the *force* option is specified.

Box 3. Appending AEDC 2018 dataset to 2009-2015 datasets

```
*Open the "smaller numeric" dataset in 09-15
use "K:\2. Data\Working datasets\AEDC 09-12-15-18\Appended\AEDC 09-12-15 smaller numeric
dataset.dta", clear

*Append the "smaller numeric" dataset in 18 to 09-15 dataset
append using "K:\2. Data\Working datasets\AEDC 09-12-15-18\Appended\AEDC 18 smaller numeric
dataset.dta", force

*Save new dataset
save "[your path here]\Analysis datasets\AEDC 09-12-15-18 smaller numeric dataset.dta", replace
```

Creating new variables and label values

After we create a working dataset, we need to clean our data and create new variables (corresponding to Do file 2. DATA CLEANING AND VARIABLE CREATION). This step involves generating new variables, defining variable labels, and defining label values. Box 4 to Box 6 show examples of these procedures.

Box 4. Example of not replacing the values in the existing variable "var27"

```
   replace var27=100 if var27>100  // do NOT do this

*Instead, you should use either "generate" or "clonevar" to create new variables:
   generate newvar27=100 if var27>100  // OR
   clonevar newvar27=100 if var27>100
```

Box 5. Example of defining label values

```
*Define "yes/no" label values
   lab def Lyesno 0 "No" 1 "Yes", replace // You can also use "modify" to replace "replace"
*Assign the above label to "Indigenous status" variable
   recode zf12m1 (1=0) (2=1) (3=1) (4=1) (.=.), gen(atsi)
   lab var atsi "Aboriginal and or Torres Strait Islander"
   lab val atsi yesno
   tab atsi
```

Box 6. Example of removing an assigned label value

```
* Remove the "yesno" label assigned to "atsi", type:
label values atsi
```

Checking variables

Each variable should always be checked to ensure the results are looking sensible. Commonly-used commands are “duplicates”, “codebook, compact”, “duplicates,” “tab”, “sum" and “sumtable” We can use these commands to check whether there are duplicates and the distribution of each variable and find out if there is an outlier.

Adding notes in do-files

There are three ways to add notes in do-files. For example, Box 7 exemplifies how to add notes when constructing a derived variable.

Box 7. Example of adding notes when creating variables

Approach 1: using "*" to add comments, for example:

```
*Recode SEP at 0-1 years into a binary variable
```

Approach 2: using "//" to add comments, for example:

```
egen earlyalltimehsas=rowtotal(bHSAs cHSAs dHSAs eHSAs fHSAs),missing  // Missing data not accounted for in this variable
```

Approach 3: using "/* … */" to add comments, for example:

```
/*
These analyses are preliminary and are based on those countries for which complete data were available by January 17, 2005.
*/
```

9.3 Appendix 3. Syntax examples of data analysis

Survey weights

The main purpose of weights in LSAC is to compensate for differences between the final sample and the national population. The weights reflect both the design of the study (to allow for unequal probabilities of inclusion in the study that may result in sampling biases) and likelihood of response (those less likely to respond are given a higher weight and those more likely to respond are given a lower weight).[31] The composition of the sample, and thus how well it represents the population, can be affected by non-participation of those chosen in the original random selection. The two main mechanisms of nonparticipation occur during the initial recruitment stage, when persons in the randomly-selected sample cannot be contacted or do not agree to participate, and during subsequent waves through attrition by loss of contact (non-contact), opting out (refusal), or otherwise moving beyond the scope of collection.[32]

The LSAC Wave 8 Weighting and Non-Response technical paper is located here: https://growingupinaustralia.gov.au/sites/default/files/tp24.pdf. Each analysis will require a different weighting variable depending on the data used. Figure 3 and 4 below detail the description of survey weights across each wave.[33] Figure 5 below lists the survey weights for LSAC Child Health CheckPoint.[34] First, choose whether population or sample weight is needed (we would usually use the sample weight); then choose the weighting variable corresponding to the waves analysed.

- Population weight: conceptually represents the number of children in the population represented by each child in the sample when creating weighted estimates. This weight would be used to produce population estimates based on the LSAC data (e.g. based on LSAC data there are approximately 22,464 infants in Australia that were never breastfed).
- Sample weight: can be used as a measure of the representativeness of each child compared to the others in the sample. This weight would be used in analyses that expect the weights to sum to the sample size rather than the population, particularly when tests of statistical significance are involved.
- Further detail on how to use the Stata survey commands are located here: https://stats.idre.ucla.edu/stata/faq/how-do-i-use-the-stata-survey-svy-commands/

In-text example

Stata 16.1 was used to conduct the analyses, with survey methods weighting to account for the probability of selecting each child in the study and non-response.[32] Because there are some primary sampling units (PSUs) in LSAC (postcodes) with only one participant, the default settings for the "svyset" command in Stata will sometimes result in standard errors being suppressed. To avoid excluding cases on this basis, standard errors for PSUs with a single observation are made by using a variance scaling factor for those PSUs so that their (within PSU) variances are equal to the average of the variances from the strata with multiple sampling units for each PSU. See example syntax below.

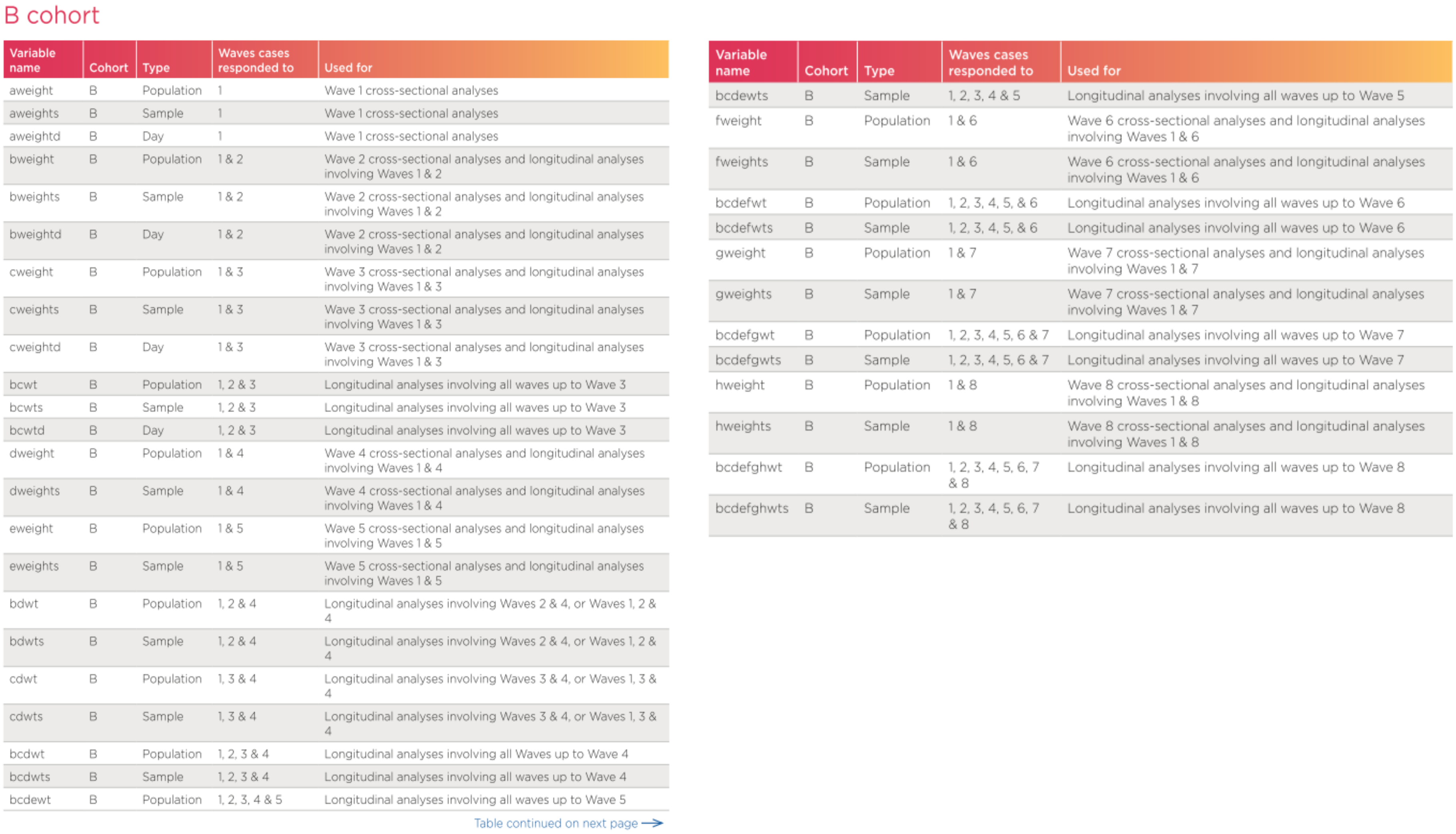
Box 8. Considering survey weights for analysis

*If the analysis is using complete data, rather than imputed data, then we use:

```
svyset pcodes [pweight=cweights], strata(stratum) singleunit(scaled)
svy, subpop(aedc_sub): logistic susch ib3.conshcn ib1.zf02m1 ib3.csep3 ib3.seifa3 ib2.cfd14a ib0.severity if anyemergSHCN==1
```

*If the analysis is using imputed data, then we often do not apply the attrition weights, instead using MI with full sample because it can account for both missingness due to attrition and item non-response:

```
mi svyset pcodes, strata(stratum) singleunit(scaled)
mi estimate, or: svy, subpop(aedc_sub): logistic susch ib3.conshcn ib1.zf02m1 ib3.csep3 ib3.seifa3 cfd14a ib0.severity if anyemergSHCN==1
```

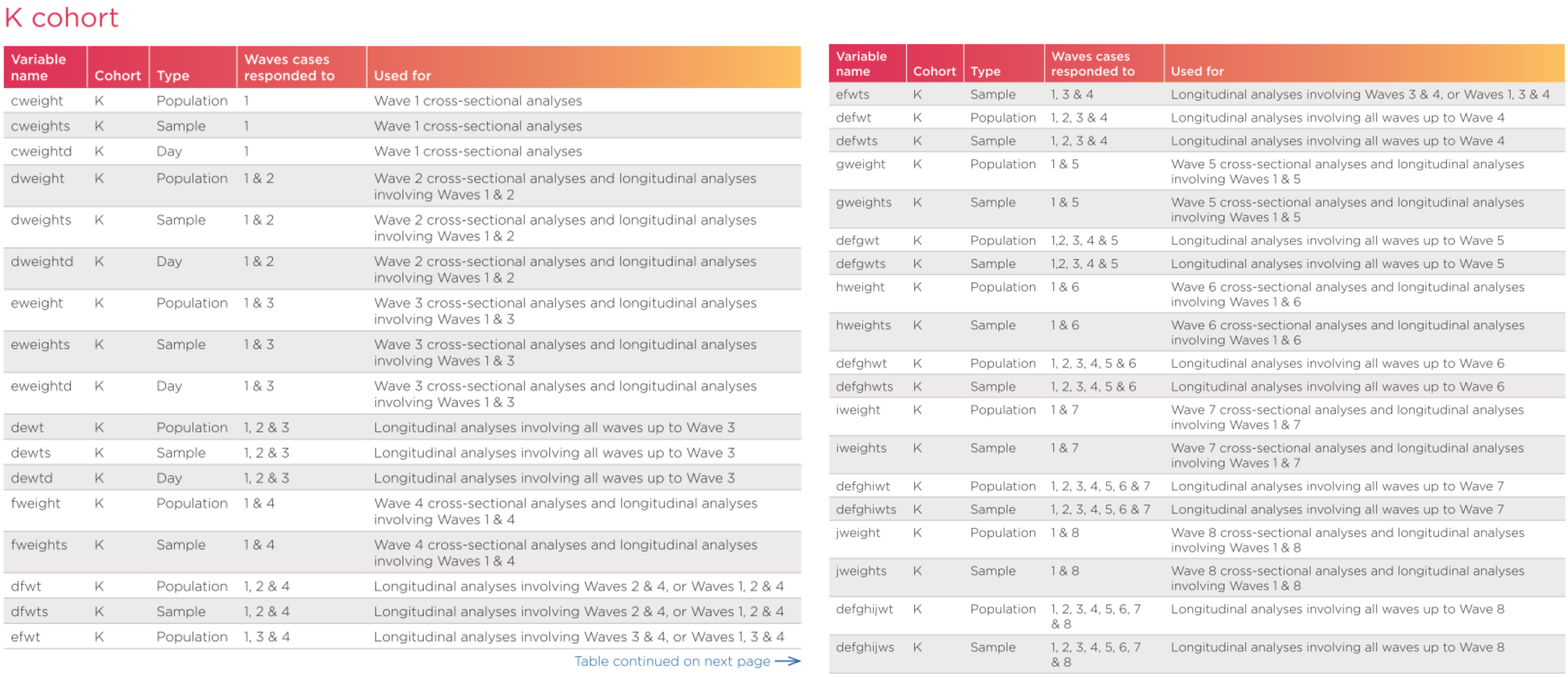


B cohort

Variable name	Cohort	Type	Waves cases responded to	Used for
aweight	B	Population	1	Wave 1 cross-sectional analyses
aweights	B	Sample	1	Wave 1 cross-sectional analyses
aweightd	B	Day	1	Wave 1 cross-sectional analyses
bweight	B	Population	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
bweights	B	Sample	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
bweightd	B	Day	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
cweight	B	Population	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
cweights	B	Sample	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
cweightd	B	Day	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
bcwt	B	Population	1, 2 & 3	Longitudinal analyses involving all waves up to Wave 3
bcwts	B	Sample	1, 2 & 3	Longitudinal analyses involving all waves up to Wave 3
bcwtd	B	Day	1, 2 & 3	Longitudinal analyses involving all waves up to Wave 3
dweight	B	Population	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
dweights	B	Sample	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
eweight	B	Population	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
eweights	B	Sample	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
bdwt	B	Population	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
bdwts	B	Sample	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
cdwt	B	Population	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
cdwts	B	Sample	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
bcdwt	B	Population	1, 2, 3 & 4	Longitudinal analyses involving all Waves up to Wave 4
bcdwts	B	Sample	1, 2, 3 & 4	Longitudinal analyses involving all waves up to Wave 4
bcdewt	B	Population	1, 2, 3, 4 & 5	Longitudinal analyses involving all waves up to Wave 5
bcdewts	B	Sample	1, 2, 3, 4 & 5	Longitudinal analyses involving all waves up to Wave 5
fweight	B	Population	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
fweights	B	Sample	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
bcdefwt	B	Population	1, 2, 3, 4, 5, & 6	Longitudinal analyses involving all waves up to Wave 6
bcdefwts	B	Sample	1, 2, 3, 4, 5, & 6	Longitudinal analyses involving all waves up to Wave 6
gweight	B	Population	1 & 7	Wave 7 cross-sectional analyses and longitudinal analyses involving Waves 1 & 7
gweights	B	Sample	1 & 7	Wave 7 cross-sectional analyses and longitudinal analyses involving Waves 1 & 7
bcdefgwt	B	Population	1, 2, 3, 4, 5, 6 & 7	Longitudinal analyses involving all waves up to Wave 7
bcdefgwts	B	Sample	1, 2, 3, 4, 5, 6 & 7	Longitudinal analyses involving all waves up to Wave 7
hweight	B	Population	1 & 8	Wave 8 cross-sectional analyses and longitudinal analyses involving Waves 1 & 8
hweights	B	Sample	1 & 8	Wave 8 cross-sectional analyses and longitudinal analyses involving Waves 1 & 8
bcdefghwt	B	Population	1, 2, 3, 4, 5, 6, 7 & 8	Longitudinal analyses involving all waves up to Wave 8
bcdefghwts	B	Sample	1, 2, 3, 4, 5, 6, 7 & 8	Longitudinal analyses involving all waves up to Wave 8

Table continued on next page →

Figure 3. Weighting variables for LSAC B-cohort (reproduced from the LSAC Data User Guide[33])

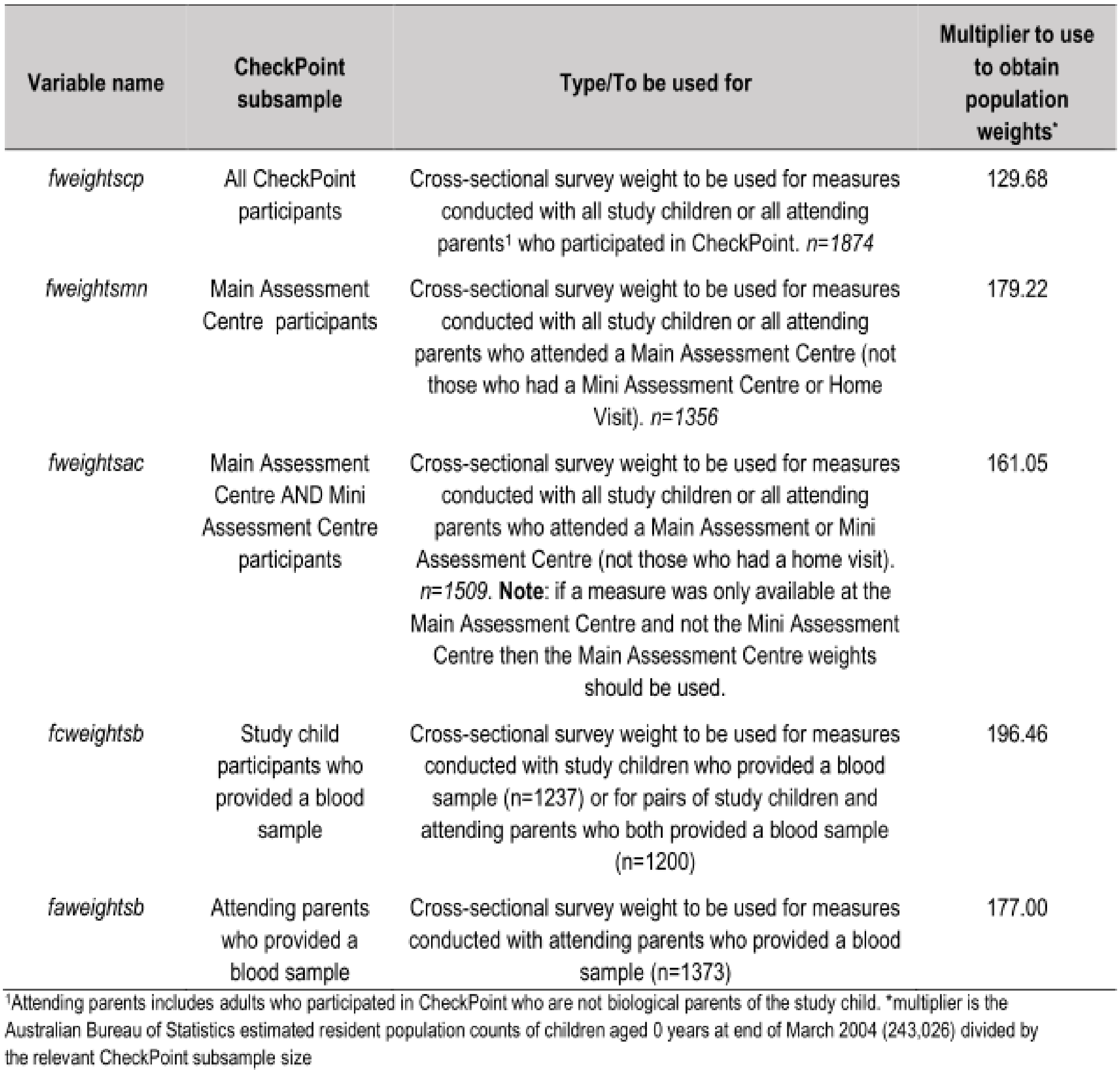


K cohort

Variable name	Cohort	Type	Waves cases responded to	Used for
cweight	K	Population	1	Wave 1 cross-sectional analyses
cweights	K	Sample	1	Wave 1 cross-sectional analyses
cweightd	K	Day	1	Wave 1 cross-sectional analyses
dweight	K	Population	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
dweights	K	Sample	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
dweightd	K	Day	1 & 2	Wave 2 cross-sectional analyses and longitudinal analyses involving Waves 1 & 2
eweight	K	Population	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
eweights	K	Sample	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
eweightd	K	Day	1 & 3	Wave 3 cross-sectional analyses and longitudinal analyses involving Waves 1 & 3
dewt	K	Population	1, 2 & 3	Longitudinal analyses involving all waves up to Wave 3
dewts	K	Sample	1, 2 & 3	Longitudinal analyses involving all waves up to Wave 3
dewtd	K	Day	1, 2 & 3	Longitudinal analyses involving all waves up to Wave 3
fweight	K	Population	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
fweights	K	Sample	1 & 4	Wave 4 cross-sectional analyses and longitudinal analyses involving Waves 1 & 4
dfwt	K	Population	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
dfwts	K	Sample	1, 2 & 4	Longitudinal analyses involving Waves 2 & 4, or Waves 1, 2 & 4
efwt	K	Population	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
efwts	K	Sample	1, 3 & 4	Longitudinal analyses involving Waves 3 & 4, or Waves 1, 3 & 4
defwt	K	Population	1, 2, 3 & 4	Longitudinal analyses involving all waves up to Wave 4
defwts	K	Sample	1, 2, 3 & 4	Longitudinal analyses involving all waves up to Wave 4
gweight	K	Population	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
gweights	K	Sample	1 & 5	Wave 5 cross-sectional analyses and longitudinal analyses involving Waves 1 & 5
defgwt	K	Population	1,2, 3, 4 & 5	Longitudinal analyses involving all waves up to Wave 5
defgwts	K	Sample	1,2, 3, 4 & 5	Longitudinal analyses involving all waves up to Wave 5
hweight	K	Population	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
hweights	K	Sample	1 & 6	Wave 6 cross-sectional analyses and longitudinal analyses involving Waves 1 & 6
defghwt	K	Population	1, 2, 3, 4, 5 & 6	Longitudinal analyses involving all waves up to Wave 6
defghwts	K	Sample	1, 2, 3, 4, 5 & 6	Longitudinal analyses involving all waves up to Wave 6
iweight	K	Population	1 & 7	Wave 7 cross-sectional analyses and longitudinal analyses involving Waves 1 & 7
iweights	K	Sample	1 & 7	Wave 7 cross-sectional analyses and longitudinal analyses involving Waves 1 & 7
defghiwt	K	Population	1, 2, 3, 4, 5, 6 & 7	Longitudinal analyses involving all waves up to Wave 7
defghiwts	K	Sample	1, 2, 3, 4, 5, 6 & 7	Longitudinal analyses involving all waves up to Wave 7
jweight	K	Population	1 & 8	Wave 8 cross-sectional analyses and longitudinal analyses involving Waves 1 & 8
jweights	K	Sample	1 & 8	Wave 8 cross-sectional analyses and longitudinal analyses involving Waves 1 & 8
defghijwt	K	Population	1, 2, 3, 4, 5, 6, 7 & 8	Longitudinal analyses involving all waves up to Wave 8
defghijws	K	Sample	1, 2, 3, 4, 5, 6, 7 & 8	Longitudinal analyses involving all waves up to Wave 8

Table continued on next page →

Figure 4. Weighting variables for LSAC K-cohort (reproduced from the LSAC Data User Guide[33])



Variable name	CheckPoint subsample	Type/To be used for	Multiplier to use to obtain population weights*
fweightscp	All CheckPoint participants	Cross-sectional survey weight to be used for measures conducted with all study children or all attending parents[1] who participated in CheckPoint. *n=1874*	129.68
fweightsmn	Main Assessment Centre participants	Cross-sectional survey weight to be used for measures conducted with all study children or all attending parents who attended a Main Assessment Centre (not those who had a Mini Assessment Centre or Home Visit). *n=1356*	179.22
fweightsac	Main Assessment Centre AND Mini Assessment Centre participants	Cross-sectional survey weight to be used for measures conducted with all study children or all attending parents who attended a Main Assessment or Mini Assessment Centre (not those who had a home visit). *n=1509*. **Note**: if a measure was only available at the Main Assessment Centre and not the Mini Assessment Centre then the Main Assessment Centre weights should be used.	161.05
fcweightsb	Study child participants who provided a blood sample	Cross-sectional survey weight to be used for measures conducted with study children who provided a blood sample (n=1237) or for pairs of study children and attending parents who both provided a blood sample (n=1200)	196.46
faweightsb	Attending parents who provided a blood sample	Cross-sectional survey weight to be used for measures conducted with attending parents who provided a blood sample (n=1373)	177.00

[1]Attending parents includes adults who participated in CheckPoint who are not biological parents of the study child. *multiplier is the Australian Bureau of Statistics estimated resident population counts of children aged 0 years at end of March 2004 (243,026) divided by the relevant CheckPoint subsample size

Figure 5. Weighting variables for LSAC Child Health CheckPoint (reproduced from the LSAC Child Health CheckPoint Data User Guide[34])

Accounting for clustered data

In the LSAC cohort, we have generally accounted for the nested nature of the data using robust standard errors, clustering on the teacher. This technique produces unbiased standard errors that allow the assumption of the independence of observations to be relaxed,[35] and produces similar results to more complex methods such as multilevel modelling that may require more robust assumptions.[36]

In-text example

In each of the following logistic regressions, we accounted for the nested nature of the data using robust standard errors, clustering on teacher. This procedure produces unbiased standard errors that allow the assumption of the independence of observations to be relaxed ,[35] and produces similar results to other methods such as multilevel modelling.[36]

Box 9. Accounting for clustered data

```
*If the analysis is using complete data, rather than imputed data, then we use:
logistic dvphys gender01, cluster(TeacherID)

*If the analysis is using imputed data, then we still account for clustering but not for attrition
weights:
mi svyset pcodes, strata(stratum) singleunit(scaled)
mi estimate, or: svy: logistic dvphys gender01, cluster(TeacherID)
```

Missing data and multiple imputations

Missing data are commonly observed during our data analysis. We will use multiple imputations to deal with missing data. According to the rule of thumb,[38] we will set up the minimum number of imputations to equal the percentage of incomplete cases. For example, 17 per cent of cases are incomplete, hence this rule would suggest 20 imputed datasets.

In-text example

The proportion of missing data across the variables was very low (an average of 3.87%). To handle missing data, multiple imputation by chained equations was conducted in Stata 16.1, producing 40 imputed datasets. The imputation model included all variables in the analysis model and three auxiliary variables (child age at time of the direct academic assessment, whether a child had repeated a grade at school, and teacher-reported academic skills at 6-7 years), as well as all two-way interactions amongst exposure and mediators. Results from each imputed dataset were combined using Rubin's rules and reported.

Box 10. Generating datasets with multiple imputation

```
*Declare MI data to be stored in the marginal long style
mi set mlong
```

```
*Identify missing values
mi describe
mi misstable summarise

*Check patterns of missing values
mi misstable patterns [var list]
mi misstable nested [var list]

*Register imputation variables (variables with some missing)
mi register imputed [var list]

*Register regular variables (variables with no missing)
mi register regular [var list]

*Impute 50 data sets
mi impute chained (regress)[continusous vars] (logit)[binary vars] (ologit)[ordinal vars]
(mlogit)[categorical vars]= [regular vars], add(50) rseed(1122) burnin(10) force augment

*Recheck whether the imputed data has missing values
mi describe
mi xeq: sum [var list]

*Check summary statistics in imputed dataset
misum [var list]

*Generate passive variables
mi passive: generate [newvar]=exp(var)
```

9.4 Appendix 4. Syntax examples of exporting results from Stata to Word/Excel

Creating tables

To avoid manual data entry errors, we can use the "*putdocx table*" command to export table results.

Box 11. Example of using "putdocx table" to export regression tables

```
*Open a dataset in Stata
sysuse nlsw88,clear
*Creating a Word document in memory
putdocx begin
*Adding a new paragraph to this active Word document
putdocx paragraph
*Adding content "EXAMPLE OF USING PUTDOCX TO EXPORT TABLES AND FIGURES" to the paragraph created by "putdocx paragraph"
putdocx text ("EXAMPLE OF USING PUTDOCX TO EXPORT TABLES AND FIGURES"), bold font("Arial", 14, blue )
*Adding a new paragraph to this active Word document
putdocx paragraph
*Adding content "Table 1. Linear regression results for wage" to the paragraph created by "putdocx paragraph"
putdocx text ("Table 1. Linear regression results for wage"), bold italic font("Arial", 13, black )

*Perform a regression on the outcome variable "wage"
regress wage ttl_exp union hours south age
*Exporting the complete regression table to this active Word document
putdocx table [tablename] = etable, border(bottom) width(100%)
*Saving this active Word document "wage.docx"
putdocx save "K:\2. Data\Data management\Analytic approaches\Example do file structure\Epi interest group\Example_Putdocx\Wage results1.docx", replace
```

Creating graphs

We can also use "putdocx image" to export STATA graphs.

Box 12. Example of using "putdocx image" to export graphs

```
sysuse nlsw88,clear
histogram wage
graph export Wage_figure.png, replace

putdocx paragraph
putdocx text ("Figure 1. Histogram of wage")
putdocx paragraph
putdocx image Wage_figure.png
putdocx save "K:\2. Data\Data management\Analytic approaches\Example do file structure\Epi interest group\Example_Putdocx\Wage results1.docx", replace
```

Melbourne Children's
A world leader in child and adolescent health

The Royal Children's Hospital Melbourne

murdoch children's research institute

THE UNIVERSITY OF MELBOURNE

Centre for Community Child Health
The Royal Children's Hospital Melbourne
50 Flemington Road Parkville Victoria 3052 Australia
www.rch.org.au/ccch

The Centre for Community Child Health is a department of The Royal Children's Hospital and a research group of Murdoch Children's Research Institute.