Avon Longitudinal Study
of Parents and Children

GUIDE TO USING ALSPAC DATA LINKED TO UK HEALTH, EDUCATION AND OTHER THIRD-PARTY DATA.

Linkage information for Research Collaborators

Version 2.0, 15.08.2023

Table of contents

Other linked datasets

ALSPAC has access to other linked datasets, including social media (Twitter), crime and spatial data (e.g. pollution, greenspace). Anyone interested in using linked data other than health or education – as detailed in this document – should contact the linkage team via alspac-linkage@bristol.ac.uk.

Overview of linked records in ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) has collected comprehensive and detailed data from ~15,000 mothers since early pregnancy, and from their children since birth. ALSPAC holds data generated from biological samples, clinical assessments, and questionnaires. These ALSPAC data can be linked to datasets from other sources, e.g., GP records, hospital statistics, education records.

Linking ALSPAC data to other datasets has several key benefits. Primarily it provides a resource-rich source of information on participants' wider lives, whilst being cost-efficient from a data collection point of view. This document provides an introduction to the types of linked health and education data available, its coverage, and how it can be requested and used. Further information on linked data (such as data dictionaries and access arrangements) can be found by contacting the ALSPAC Linkage team at alspac-linkage@bristol.ac.uk. Researchers planning to use ALSPAC linked data should read this document in conjunction with the most recent version of the ALSPAC Access Policy.

Linked datasets must be accessed through the ALSPAC Data Safe Haven, an ISO27001 accredited resource for secure record linkage and analysis. The Data Safe Haven comprises the Linkage local secure server in Bristol and the Secure eResearch Platform (UKSeRP) server, leased from Swansea University. Disclosure control, and data flow into and out of UKSeRP, is controlled by ALSPAC (see Sections 4.2 and 4.4 for details).

There are three generations of ALPSAC participants: the original mothers and their partners (referred to as Generation 0, or G0s); the children of the 90s, known as the index participants (Generation 1, or G1s); and the children of the children of the 90s (G2s). It should be noted that the majority of Linkage data is available only for G1s. Information on the cohort profiles is available at: bristol.ac.uk/alspac/researchers/cohort-profile/.

Table 1 summarises the datasets with current permissions to link to ALSPAC participant data, and with approximate cohort sizes and data availability. ALSPAC has a comprehensive consent process, enabling participants to opt-out of data linkage (and other aspects of data collection and usage) at any point. More information is available here: bristol.ac.uk/alspac/participants/using-your-records/

Table 1. Availability of linked health and education data for ALSPAC participants

Source Data	Coverage: years (Approximate age of participants)	Maximum sample size [1]	Availability [2]
HEALTH – national records			
Primary Care (GP)	1990 – 2016 (Ages 0 – 25)	12,000 (G1s)	Via UKSeRP for projects based at a UK Research Institution [2]
Hospital Episode Statistics (HES): admitted patients	1990 – 2017 (Ages 0 – 27)	11,000 (G1s)	Via UKSeRP for projects based at a UK Research Institution. At least one senior member of the research team must be based at, or have an honorary contract with, the University of Bristol.
Hospital Episode Statistics (HES): outpatients	2003 – 2017 (Ages 13 – 27)	10,000 (G1s)	
Hospital Episode Statistics (HES): A & E	2007 – 2017 (Ages 17 – 27)	9,000 (G1s)	
Mental Health Services Data Set (MHSDS)	2006 – 2015 Ages (16 - 25)	800 (G1s)	
HEALTH – local records			
STORK (midwifery and delivery records of G0s & G1s)	1990 – 1992 (Age 0)	11,500 (G0s) 11,500 (G1s)	Via UKSeRP for projects based at a UK Research Institution [2]
Avon & Wiltshire mental health partnership – community mental health care	TBC	1,200 (G1s)	
HEALTH – specialised datasets			
Congenital abnormalities (from the Avon Child Health System)	1990 – 1992 (Age 0)	240 (G1s)	Direct from ALSPAC
AvonCAP: Bristol-based study into Community Acquired Pneumonia	2020 – 2022 (various)	120 (G0s) 50 (G1s)	Via UKSeRP for projects based at a UK Research Institution [2]
Bristol Self-Harm Register (BSHR)	2010 – 2018 Ages (17 – 27)	160 (G1s)	
EDUCATION			
National Pupil Database (NPD) Key Stage (1-5)	1995 – 2011 (Ages 5 – 18)	13,000 (G1s)	Via UKSeRP for projects based at a UK Research Institution [2]
Absences & exclusions	2006 – 2009 (Ages 15-16)	11,000 (G1s)	
Annual School Census	1999 – 2009 (Ages 8 – 18)	20,000 (G1 schools)	

NOTES

[1] Sample sizes: Due to missing data and changes in consent status, actual sample size will vary by time point and variable, and can be considerably smaller than the maximum sample sizes shown here.

[2] Data availability: For access to these datasets, at least one member of the research team must be based at, or hold an honorary contract with, a UK Research Institution.

Accessing linked records

This document outlines the typical series of events stemming from a request to use linked data. The process of applying for, and using, linked data can be summarised as five distinct phases:

1. Project Setup

2. Data definition and verification

3. Confirmation of project team

4. File building and data release

5. Account close-down and archiving

Each of these is outlined below. The processes differ slightly depending on whether the request is to access health or education data, and these are addressed separately in Phase 2 (Data definition and verification). Anyone intending to use more than one form of linked data should familiarise themselves with both sections.

Phase 1: Project Setup

1.1 Submitting a proposal via the ALSPAC Online Proposal System (OPS)

Before deciding to submit a research proposal, researchers should familiarise themselves with ALSPAC's guide to accessing data: http://www.bristol.ac.uk/alspac/researchers/access/. A proposal should detail project logistics, aims & objectives, methods, and data sought. The use of linked records within the context of the project must be justified, and only data directly relevant to the research question should be requested. Particular attention should be paid to justifying the use of sensitive variables (e.g., mental/sexual health, abuse, termination of pregnancy, free school meals, special educational needs). Researchers should contact alspac-linkage@bristol.ac.uk if they need any additional information about third party data before submitting a proposal.

1.2 OPS approval, costing and conditions

The proposal will initially be considered by the ALSPAC Executive. Depending on the nature of the linked data requested, the Linkage team may arrange a meeting with the researcher to discuss the project in more detail and to provide some preliminary information on likely timescales, feasibility, and cohort sizes. If, after discussion, a proposal to use linked data is approved, the costings and conditions of linked data access will be communicated to you by the Executive and a data buddy will be allocated (see ALSPAC Access Policy). If the proposal is not approved, further discussions with the ALSPAC Executive or the Linkage team may be able to improve the likelihood of a subsequent submission being approved. Information about accessing ALSPAC data more generally is available in the ALSPAC Access Policy. Illustrative costs for linkage data are outlined in Appendix B.

1.3 Additional approvals for linked health data

Once approved by the ALSPAC Executive, projects proposing to use linked health data are submitted as amendments either to the ALSPAC Ethics and Law Committee (ALEC) or the NHS Research Ethics Committee (REC), or in some cases possibly both, for final approval. The approvals required for a project will depend on the specifics of the project and the data required; ALSPAC can assist with guidance on this. For particularly sensitive information (e.g., mental and/or sexual health, termination of pregnancy, abuse) requested from secondary care (HES) records, an additional application will need to be submitted to the Health Research Authority (HRA) Confidentiality Advisory Group (CAG). See Section 2.a2 for further details.

Please note the individual approvals (ALEC, REC, CAG) occur consecutively, and can each take up to three months to be approved.

1.4 Project lay summary for website & social media

As part of ALSPAC's fair processing procedure, a lay summary is required to notify participants of projects that propose to use linked records. The lay summary should include a full summary for the ALSPAC website, and a shortened version (280 characters max.) to be used in social media posts. The lay summary must be on the website for four weeks before any data can be provided, to enable ALSPAC participants the opportunity to opt-out of specific projects. Please see the ALSPAC website and social media accounts for examples.

Phase 2: Data definition and verification

All requested linkage variables must be directly relevant to the research question outlined in the original proposal (Phase 1). Data Request forms must be completed, and approved, for both linked data and ALSPAC-collected data before a project can proceed. If the scope of the project changes, an amendment must be submitted via the Online Proposal System (OPS), and further charges may be incurred.

It should be noted that omics datasets are usually too large to be directly linked to health or education data in UKSeRP. Researchers using omics data are advised to derive aggregate variables which can then be uploaded to UKSeRP.

The processes for requesting and accessing health and education data are addressed separately below.

2a: Requesting linked health data

2a.1 Define data request

Based on project setup and conditions communicated in Section 1.3, the researcher should define a list of medical codes specific to the research question. The coding system required depends on the dataset and timepoint(s) of interest. For example, Hospital Episode Statistics (HES) diagnosis codes are coded to ICD-9 (pre-1995) and ICD-10 (post-1995). GP diagnoses and prescription data use both the Read Code and the SNOMED code systems. The Linkage team can advise on the coding system(s) relevant to each project. These links provide an overview of the coding systems:

- Read codes: https://isd.digital.nhs.uk/trud3/user/guest/group/0/pack/9
- ICD9/10 codes: http://www.icd9data.com/ and http://www.icd10data.com/
- SNOMED: https://termbrowser.nhs.uk/

The website https://clinicalcodes.rss.mhs.man.ac.uk/ provides a repository of code lists linked to publications that have used electronic medical records. It may also be possible to identify linked health records of interest by free-text matching to code descriptors. This approach is unlikely to be appropriate for all projects, so should be discussed with the Linkage team before being pursued.

It should be noted that the age profile of the ALSPAC index participants, and the dates covered by linked health data (Table 1), mean that many health conditions occur less frequently in the ALSPAC G1 cohort than they do in the general population. The Linkage team may be able to provide information on the approximate incidence of individual medical codes within the cohort. Researchers are advised to consult the Linkage team before finalising code lists. (More information on how the Linkage team can address low numbers of cases, or small cell counts, are dealt with in Section 4.2 on disclosure control.)

2a.2 Ethics

Researchers requesting linked health data must provide a description of the project to be added to the Linkage team's list of approved projects. This description is submitted as an amendment to the ALSPAC Ethics and Law Committee (ALEC) and to the NHS Research Ethics Committee (REC) on the researcher's behalf. All project descriptions should include:

1. Overview of the research to be conducted, with 2-3 key references
2. Demonstrated need for the research, emphasising and referencing calls for evidence (particularly from the NHS)
3. Justification for why ALSPAC is an appropriate setting
4. Hypotheses, data sources, and statistical methods

The Linkage team can provide example templates for the project description and, once complete, will manage the approval process with the ALEC and the REC.

If the intended exposure/outcome of interest is deemed particularly sensitive (e.g., mental and/or sexual health, termination of pregnancy, abuse), and is part of secondary care (HES) data, an additional application will need to be submitted to the Health Research Authority (HRA) Confidentiality Advisory Group (CAG). Decisions from the ethics review will be fed back to the Linkage team and, if approved, the project can proceed.

2b: Requesting linked education data

The Linkage team has access to variables from the National Pupil Database (NPD), including Key Stage (1-5), Pupil Level Annual School Census (PLASC), and Annual School Census (ASC). Data on exclusions & absences, and English as an Additional Language are also available. A full list is available from the Department for Education, and is also provided in the Linkage variables catalogue (available on request from alspac-linkage@bristol.ac.uk).

It should be noted that the use of linked education records can only be approved if the intended research outputs will help promote the education or well-being of children.

Variables can be requested by completing the relevant tabs in the Linkage Data Request form, ensuring that it matches the approved proposal. Note that some variables, e.g., special educational needs (SEN), will require additional justification in the proposal.

Phase 3: Confirmation of project team

3.1 Data Access Agreement (DAA)

The ALSPAC Data Access Agreement (DAA) must be signed before any data can be accessed, including data accessed via UKSeRP. This is a legal document laying out the terms and conditions under which ALSPAC shares data and cannot be changed. An ALSPAC data buddy will prepare a project specific DAA once both the linkage and non-linkage data requests have been finalised. Researchers are sent a copy of any Data Sharing Agreements/Contracts associated with the data they propose to access. Please note that access to an entire linked dataset will be governed by the controls needed for the most sensitive component (i.e., a project using both mental health and education data will be subject to the restrictions applied to the mental health data.)

3.2 Data User Responsibilities Agreement (DURA)

All researchers who will access the data are required to complete a DURA. A sample can found here. The ALSPAC data buddy will provide a link to the online form. Once the DURA and DAA are completed and fees have been paid, the data building process can begin (Section 4.1). Please see Appendix B for current cost estimates.

3.3 Training

Researchers accessing linked health or linked education data within the secure research environment (UKSeRP) will need to complete at least **one** of the following training courses:

1. ONS Approved/Accredited Researcher status, details available here.
2. Safe User of Research data Environments (SURE) Training course – run by ONS, the UK Data Service, the Administrative Data Research Network.
3. MRC – Research, GDPR and confidentiality training, available online.

Proof of completion of one of these courses is required before accessing UKSeRP. This can be provided by sending a screenshot or pdf of the course certificate to the Linkage team. Researchers will be required to provide evidence of subsequent training if their project's planned end date is after the training certificate's expiration date.

Phase 4: File building and data release

4.1 Data building process

a) An ALSPAC data buddy creates a dataset of ALSPAC-collected data (questionnaires, clinical measures, etc) based on the finalised data request form. This dataset is securely transferred to the Linkage team who manage its further transfer into UKSeRP.

b) In UKSeRP, the Linkage team links the dataset with health and/or education data, as defined in the project setup (Phase 1)

c) Most linkage data is provided in 'long' format (i.e., more than one line of data per person) as there is likely to be more than one recorded 'event' per person, e.g., GP visits or school absences. The provision of data in alternative formats can be discussed with the Linkage team prior to the data being released.

d) Data are de-identified, and disclosure risk assessed and controlled. See Section 4.2 for further information on minimising disclosure risks.

e) Once UKSeRP accounts are setup and permissions configured, login details are shared with the researcher(s).

4.2 Pre-release disclosure control

All ALSPAC data are curated to maintain absolute confidentiality of participants. To ensure this, standard procedures are applied to all linked datasets:

a) **Dates** – all dates are aggregated to the nearest week, month, or year, as appropriate for the circumstances. Researchers are never supplied with full dates of birth, or dates of events (e.g., GP visits).
b) **Small cell counts** – all data will be assessed for rare groups. The benchmark for rarity is n=5, i.e., any cell counts with <5 participants in a final dataset will be suppressed or, in consultation with the researcher, aggregated into larger groups.
c) **Health data** – are minimised to the defined code list (Section 2a.1). If codes are recorded for <5 participants, codes are grouped, suppressed or recoded to binary variables to avoid small cell counts.
d) **Education data** – are minimised to the requested variables (Section 2b) and, if any variables enable the identification of <5 participants, are further grouped, suppressed or recoded to avoid small cell counts.

The above disclosure controls also apply to any dataset provided by a researcher if linkage to ALSPAC data makes it disclosive. Other forms of disclosure control may also need to be implemented prior to data release; these will be discussed with the research team to maximise data availability while minimising disclosure risk.

4.3 Analyses

Analyses take place in UKSeRP, in a Windows environment, using the software package (e.g., Stata, SPSS, R) defined at the project setup (Phase 1). Other software packages are available on request but may be subject to additional licence costs. N.B. Access to the internet is not possible from inside UKSeRP.

Files generated outside UKSeRP (e.g. scripts, code lists, reporting templates) can be imported via a 'File-In' request available on the UKSeRP portal. 'File-Ins' must not contain individual-level data or information which could potentially re-identify any ALSPAC participants. 'File-Ins' are checked by the Linkage team prior to being made available to researchers in UKSeRP.

Researchers cannot copy or move data out of UKSeRP. Any output items (e.g., tables, graphs, summary statistics) to be retrieved from UKSeRP must be dropped into the 'File-Out' area available on the UKSeRP desktop (see Section 4.4).

Any publications based on ALSPAC data need to adhere to the ALSPAC publications checklist and should be submitted to the ALSPAC Executive Committee for approval prior to submission for publication.

4.4 Retrieving analysis outputs from UKSeRP

Analysis outputs can be retrieved from UKSeRP via the 'File-Out' folder on the platform's desktop. 'File-Out' items will be disclosure checked by a member of the Linkage team. Researchers receive a notification when files pass disclosure control, and a weblink to download the files. Files that do not meet disclosure control criteria will be returned with a note detailing the issues.

Small cell counts: no tables or graphs included in a 'File-Out' request should contain cell counts of less than 5 (including zero). Researchers should collapse categories, or replace the cell counts with '<5' to mitigate small cell counts. If the cell contains zero, a footnote should be included to clarify that '<5 may also include zero'. This also applies to any imputed data. Percentages associated with absolute numbers must be dealt with in a similar manner.

The initial data access 'cost recovery' charge covers two bulk output disclosure checks a year. Each check can contain multiple items (e.g., summary tables, graphs). Additional output checks may incur further charges.

Phase 5: Account close-down and archiving

Researchers should notify the Linkage team when they have completed their analyses and/or no longer require UKSeRP access. When analyses are complete, the project data will be archived, and access to the data removed.

Copies of any variables derived from the original data, along with appropriate documentation, should be returned to ALSPAC at the end of the project. The process for this is outlined in the ALSPAC Access Policy.

The Linkage team will contact researchers for a status update if the project's access to UKSeRP is due to expire before analyses are complete. Researchers can apply for an extension of access via an amendment to the proposal using the online proposal system. Please note there are additional charges for extending access to UKSeRP.

Appendices

A: Timescales

The time taken from submitting a research proposal to the linked data being available to the researcher varies enormously. For projects where funding is already secured, personnel are in place, and the variables of interest are identified from the start, dataset creation can start as soon as all the necessary agreements are in place and payment is made (see Phase 3). For projects waiting for funding, or where there's uncertainty about the variables of interest, it can take much longer to get to this point. Once the data request has been finalised (see Sections 2a.1 and 2b.1), all the necessary agreements are in place and payment made, it can take up to three months for the ALSPAC-collected data to be transferred to UKSeRP, matched with Linked data, quality assured, and disclosure controlled.

B: Costs

Table 2 and Figure 1 (over page) outline data access costs for projects seeking to use linked health or education via UKSeRP. Please note, these are indicative costs and are meant as a guide to be used for project-planning and grant applications. Exact costs will be confirmed by the Linkage team as part of the project approval process (see Sections 1.2 and 1.3).

Table 2. Data access costs for linkage data in UKSeRP

Item		Details	Cost [1]
UKSeRP costs			
Year 1	Base cost	Required by all projects. Covers access to UKSeRP for one year	£1,100
	User cost	All projects require at least one user who must be named on the project proposal (or subsequent amendment)	£550
	Disclosure control	All projects require this. It covers two bulk output disclosure checks. Additional output checks will be charged at £660 [2]	£660
Years 2+	Base cost	Required by all projects using UKSeRP for more than 12 months. Covers one additional year of access.	£550
	User cost	Required by all projects using UKSeRP for more than 12 months. Covers one additional year of access.	£275
	Disclosure control	All projects require this. It covers two bulk output disclosure checks. Additional output checks will be charged at £660 [2]	£660
Ethical approval			
NHS Research Ethics Committee (REC) amendment		Required for all projects using linked health records	£400
Health Research Authority (HRA) Confidentiality Advisory Group (CAG) application		Required for projects seeking to use linked secondary care health records (HES, MHSDS) where the outcome or exposure is particularly sensitive (e.g., mental health, sexual health)	£1,320
File building			
Education data		See Sections 2b and 4.2	£660
GP data		See Sections 2a and 4.2	from £660
HES data		See Sections 2a and 4.2	from £660
MHSDS data		See Sections 2a and 4.2	from £660

NOTES

[1] Costs: Correct at the time of writing (August 2023). All costs are subject to VAT where appropriate. These costs are for accessing linked data in UKSeRP and are in addition to the access fee for ALSPAC-collected data. More information is available in the ALSPAC Access Policy.

[2] Disclosure control: Each check can contain multiple items (e.g., summary tables, graphs)

Figure 1. Costing examples

Example 1: Access to linked education data for one user for one year

Item	**Cost [1]**
UKSeRP base cost	£1,100
User cost	£550
Disclosure control	£660
Building education data	£660
Total cost	**£2,970**

Example 2: Access to linked education and health (HES) data for one user for one year, with mental health secondary care data as an outcome or exposure. (HES data required only at the at highest, alphabetical, level of ICD-10 codes (see Section 2a.1).)

Item	**Cost [1]**
UKSeRP base cost	£1,100
User cost	£550
Disclosure control	£660
NHS REC amendment	£400
HRA CAG application	£1,320
Building education data	£660
Building HES data	£660
Total cost	**£5,350**

Example 3: Access to linked health (GP and HES) data two users for two years. (GP Read Codes and SNOMED code data required as both exact and partial code matches, plus text matching of descriptive fields. HES data matched on both ICD-9 and ICD-10 codes, to two characters, plus corresponding text fields. See Section 2a.1)

Item	**Cost [1]**
UKSeRP base cost & 1 user (year 1)	£1,650
User costs (2nd user, year 1)	£550
UKSeRP base cost & 1 user (year 2)	£825
User costs (2nd user, year 2)	£275
Disclosure control (year 1)	£660
Disclosure control (year 2)	£660
NHS REC amendment	£400
Building GP data	£1,980
Building HES data	£1,320
Total cost	**£8,320**

NOTES

[1] Costs: correct at the time of writing (August 2023). All costs are subject to VAT where appropriate. These costs are for accessing linked data in UKSeRP and are in addition to the access fee for ALSPAC-collected data. More information is available in the ALSPAC Access Policy.

C: Further information

Please contact the Linkage team who will be able to answer any questions about accessing ALSPAC participants' linked records:

E: alspac-linkage@bristol.ac.uk

W: https://www.bristol.ac.uk/alspac/researchers/our-data/linkage/

T: @CO90s

D: Change history

This appendix details the changes made to this document since the release of v1.0 in March 2023.

v1.1 Released 16.05.2023

Sentence added to Section 2.b outlining that 'linked education records can only be approved if the intended research outputs will help promote the education or well-being of children'

Explanation of 'File-Out' process expanded in Section 4.3

v2.0 Released 15.08.2023

Reference to COVID-19 data removed in Table 1

Section 2a.1 expanded to include the option of free-text matching of medical codes

Removal of all references to Linkage costs being in the Data Access Policy

The sections 'Timescales' and 'Further Information' moved into an Appendix

Costs added as Appendix B

Change history added as Appendix D

* * * * * * *