ALSPAC DATA MANAGEMENT PLAN, 2019-2024

0. Proposal name

The Avon Longitudinal Study of Parents and Children (ALSPAC). Core Program Support 2019-2024.

1. Description of the data

1.1 Type of study

ALSPAC is a multi-generation, geographically based cohort study following 14,541 mothers recruited during pregnancy (in 1990-1992) and their partners (G0), offspring (G1) and grandchildren (G2).

1.2 Types of data

Quantitative data

- Data from numerous self-completed paper-based/online **questionnaires**.
- Data from **clinic-based assessments**: physiological, cognitive and anthropometric measures, structured interview data and computer based questionnaire data.
- Genetic, metabolomic, proteomic, epigenetic, biochemical and environmental exposure data obtained from analysis of **biological samples**.
- Data derived from **images** collected as part of clinical assessments (including MRIs, Liver scans, DXA, PqCT, retinal scan, 3D Face and body shape).
- Data obtained through **linkage to administrative records** including maternity and birth records, child health records, cancer/death registrations through ONS, Primary and Secondary health care records and education and criminal Records.
- Data obtained from social media including Twitter, Facebook and Instagram

Qualitative data from sub studies

- Small sub studies involving direct interview of participants or focus groups; audio/transcript files being generated.

Bio resource

- Biological samples collection including DNA, lymphoblastic cell lines (LCLs), blood, saliva, urine, hair, tissue such as placenta and umbilical cord, teeth and nail clippings**.**

1.3 Format and scale of the data

Type and scale of data

- 14,541 mothers originally enrolled producing 14,676 fetuses with 13988 G1 children still alive at one year of age. Additional waves of enrollment resulted in a further 913 G1 children joining the study since 2000.
- At the time of writing (Feb 2019), > 900 G2 children (including *in utero*) have enrolled and provided data.
- Self-completion questionnaires are electronically captured (scanned or digitally collected); to date there have been 48 questionnaires completed by mothers (n=6000-13500), 18 by partners (n=3000-9500), and 34 by G1 (n=5000-8000). Currently, 14 different questionnaires are used to collect data about G2 children.
- Data from clinical examinations are electronically captured (scanned paperwork or digitally collected). There have been 4 maternal sweeps (n=3500-4700), 1 father sweep (n=2000) and 10 G1 sweeps (with a minimum of 4000 attendees at each sweep).
- Electronically captured (scanned from paper or digitally collected) enrolment and consent (for e.g. record linkage, obtaining and using biological samples) forms.
- Image data e.g. face shape, DXA, MRI, liver scan, 3D body scan.
- Administrative data (e.g. educational records in the National Pupil Database, linkage to the NHS central register) The scale of linkage to individual administrative data sources depends on the completeness of these sources, the relevant participant consents and other permissions alongside technical considerations relating to the systems where these data are held.
- Data obtained from biological samples including whole genome sequence data (n=2000), genome wide association (GWAS) data (n~22,000), metabolomics data (from ~20,000 samples), genome wide epigenetics data (n~5000) plus small scale bespoke biochemical, cellular and genetic analysis.

- Bio resource: over 30,000 DNA samples, 15,000 LCLs and over 1 million sample aliquots (blood, urine, saliva, tissue, hair, nails).

Data formats

- ALSPAC stores raw numerical data in a number of formats depending on the source of the data, including MS Access, SQL Server databases, MS Excel compatible files (.csv and .xls format), REDCap (MySQL), SPSS save and portable formats and Stata data files. Text data are stored as flat text files, MS Access databases and .csv files.
- Molecular data are stored in flat file formats including .csv and JSON allowing for more complex file structures. Some formats are program specific (e.g. PLINK for SNP data). Some data are stored in MySQL databases (e.g. methylation data) but some molecular data are unsuitable for databases due to increasingly long index lengths; working copies of these 'Big Data' and archived data are stored on University of Bristol (UoB) storage systems (e.g. the ACRC Research Data Storage Facility). Raw (laboratory) data (e.g. Illumina IDAT format genotyping/ methylation files) will also be redundantly archived on UoB storage systems ensuring future availability.
- Data made available for research use through the ALSPAC resource is stored internally in multiple formats. Data is curated using a statistical package and stored in both SPSS and Stata formats. The curated data is imported into a Data Warehouse Opal/MongoDB) and is stored in a highly flexible structure. Custom datasets can be exported in any of the common statistical formats, including SPSS, STATA, SAS, R, csv (Excel).
- Multimedia data such as DXA scans, recorded participant interviews and face shape images are stored using uncompressed or lossless compression formats where possible.
- Where applicable data formats may be migrated as new technologies become available and are proved robust enough to ensure digital continuity and continued availability of data.

2. Data collection / generation

2.1 Methodologies for data collection / generation

New data will be generated by:

- Questionnaires (online or paper-based) completed by all cohort groups.
- Clinic based assessments on all cohort groups.
- Further biological sample collection from all cohort groups
- Linkage to administrative records of G0, G1 and G2.
- Interviews from qualitative studies.
- Image files such as DXA scans on G0 and G1.
- Updated contact information as provided by study participants.
- Molecular laboratory analysis of existing and new samples and integration with public molecular data.
- Further biological sample collection (blood, urine, saliva, hair, placenta, umbilical cord, breast milk, meconium, stools, DNA and cell lines) from all cohort groups.
- Social media, such as Twitter.

2.2 Data quality and standards

- Each data item to be assessed by logical and range checks built into electronic data collection systems, with ambiguous values assessed by an operator.
- All assessment scales used in either questionnaires or clinic-based assessment to have been validated externally with a known reference paper.
- A small sample (~3%) of clinic participants to be re-invited to the clinic to validate earlier measures and test for any possible fieldworker bias or equipment calibration issues.
- Interview data to be collected and validated in real time on encrypted laptops with data routinely transferred to the central repository.
- Clinical assessment data to be collected by trained fieldworkers according to clear protocols; data to be analysed regularly by research staff to ensure data quality standards are being met; regular audits of clinic processes will be performed.
- Repeat molecular analysis of a subset of samples, QC using control probes and analysis for batch effects (built into the ALSAPC LIMS [laboratory information management system]).
- The laboratory managing the bio resource has obtained the ISO9001 quality standard. Sample data are stored in ALSPAC LIMS.

3. Data management, documentation and curation

3.1 Managing, storing and curating data.

- Data from all sources will be cleaned, and prepared for analysis by the in-house statistics, bioinformatics and data management teams following established standard operating procedures (SOPs).
- Each data item is referenced and stored using a universal indexing and naming convention.
- Research data items are stored separately from administrative data (including subject identifiers) and linked through anonymised files accessible only to specified members of the data management team.
- Instrument specific and bespoke data formats are archived 'as is' to ensure the integrity of the original source material. As with other data sources, original copies of the data are never altered in any way.
- All research data collected are catalogued, maintained and archived on University of Bristol infrastructure which is scalable, secure and backed-up routinely.
- The Bio resource is licensed by the Human Tissue Authority (license number 12512) and samples are stored in secure freezers and cryostores. These facilities are linked to an emergency generator to provide back-up power and are covered by a 24hr alarm system to alert staff of freezer failures out of normal working hours. The cell line collection is backed up at an external second site.

3.2 Metadata standards and data documentation

Metadata are collected as an integral process to (i) catalogue and index the data in a searchable manner, (ii) define the assessment tools (validated measures, key reference publication, modifications etc), (iii) describe the data collection process on an individual basis (age at completion, administration and reminder process) (iv) catalogue laboratory information as captured through LIMS and (v) assign a geographical reference point (at a non-disclosive level) to assist spatial analysis.

- Research data are not made available until it has been fully documented and published on the ALSPAC website (http://www.bristol.ac.uk/alspac/researchers/our-data/).
- Metadata is provided through CLOSER Discovery (https://discovery.closer.ac.uk/) and is compatible with the Data Documentation Initiative (DDI) Life-cycle 3.2.
- Study protocols, assessment tools, data derivation methods and coding schema are provided as part of the research data documentation available as downloadable content from the ALSPAC website (http://www.bristol.ac.uk/alspac/researchers/our-data/).

3.3 Data preservation strategy and standards

- It is envisaged that ALSPAC will continue to operate as a resource for current and future generations indefinitely. In line with this expectation all data are anticipated to be maintained indefinitely.
- ALSPAC maintains an archive of data which is available to researchers on request and maintained on secure servers which are backed up on a regular basis. The University infrastructure consists of real time mirroring of data across two geographically separate data centers (Bristol and Slough) as well as off-site tape backups on a nightly basis.
- In order to ensure the longevity and availability of the resource, ALSPAC reviews data on a regular basis and will apply digital continuity methods where applicable to migrate data formats at risk of obsolescence to newer formats to ensure that information contained within the files remains complete and usable.
- Where data are disposed of (for example data that have been secured elsewhere on an obsolete hard disk) this will be done securely and in line with University IT Information security policies.
- Primary source material (E.g. Questionnaires, clinic data sheets and consent forms) will be preserved as electronic (scanned) copies where practicable.
- The UoB library holds an extensive administrative archive (Special Collections reference DM2616) of catalogued paperwork up to 2005 (study grant applications, protocols, ethical approvals, participant information materials, keying and coding specifications, documentation on each data collection measure and their provenance and the file building syntax). https://www.bristol.ac.uk/library/special-collections/strengths/alspac/.

4. Data security and confidentiality of potentially disclosive information

4.1 Formal information/data security standards

- ALSPAC gained ISO 27001 certification in 2012 and is compliant with that standard

4.2 Main risks to data security

ALSPAC avoids potentially disclosive subsets of data at all costs. For example:

- Complete dates (of birth, Q completion, clinic attendance) are not released to researchers; instead

ages are derived.

- Cell counts are no smaller than n=5; this applies to all publications that are reviewed by the ALSPAC Executive prior to journal submission.
- Free text (with its own unique ID) is coded separately from any other data; any identifying information is screened out before being passed to a researcher.
- Address data is dealt with separately to any other data (with its own unique ID), relevant address level data are obtained, aggregated as appropriate and matched back to the main dataset.
- All external researchers receive a dataset with an ID attached which is unique to them and their project.
- Anonymised frequency tables and summary statistics are freely available as part of the data dictionary.

5. Data sharing and access

Identify any data repository (-ies) that are, or will be, entrusted with storing, curating and/or sharing data from your study, where they exist for particular disciplinary domains or data types. Information on repositories is available here.

5.1 Suitability for sharing

Yes. ALSPAC data is used in a wide and varying number of research consortia and cross-cohort collaborations.

5.2 Discovery by potential users of the research data

- The ALSPAC access policy details how data can be accessed by researchers (http://www.bristol.ac.uk/alspac/researchers/data-access/).
- The cohort is advertised through a wide variety of sources including the MRC gateway to Research and the Maelstrom Catalogue,
- ALSPAC is the most commonly searched study in CLOSER Discovery.
- The ALSPAC web site describes the cohort and available data (http://www.bristol.ac.uk/alspac/researchers/our-data/), and hosts data documentation and catalogues. The data dictionary is fully searchable by keyword.
- A bespoke variable search tool is available (http://variables.alspac.bris.ac.uk). This enables a quick search of the data and facility to download a list of variables. Within the next twelve months this will be replaced with the Mica web portal (https://www.obiba.org/pages/products/mica) allowing for advanced searching of variables and creation of a variable list for submission as a data access request

5.3 Governance of access

ALSPAC is committed to providing access to ALSPAC data to the widest possible research community. Currently, ALSPAC data are made available to researchers on a supported basis rather than via an unrestricted, open resource. Bespoke datasets of requested variables are provided to collaborators by a data preparation and statistics team upon completion of a Data Access Agreement (http://www.bristol.ac.uk/alspac/researchers/access/)
Briefly, available data are described on the ALSPAC website. Researchers wishing to access these data submit a proposal to the ALSPAC Executive committee. Approval for access is given if the data requested are available and their release does not (i) risk disclosure of participant identity; (ii) violate any ethico-legal or other stipulations that apply to ALSPAC; or (iii) run the risk of harming the study as a whole or any participants in it. Crucially, data are viewed as a non-finite resource and proposals for access are therefore *not* subject to formal scientific review. Once an application has been awarded, specified variables are then provided to the investigator by a 'data buddy', who supports the user with data descriptors and additional variables as required.

Requests for data acquired via linkage to routine health and administrative records are subject to access constraints determined both by ALSPAC and the original data owner. These constraints can involve seeking additional project clearances with the original data owner, the statistical modification of the data to control for disclosure risks and ethical approval (for requests involving health data). All linkage data are filtered for participant consent at the time of release and this can impact on the resulting sample size. Access conditions are subject to changes applied by the data owner which are outside ALSPACs control. All access requirements must be adhered to at all stages of the research cycle. Researchers accessing these data do so under a legally binding contract.

Researchers will access linked health records via the UK Secure eResearch Platform (UKSeRP), Swansea University. This will be managed by the ALSPAC data linkage team who control access

permissions and researcher outputs. The system will allow researchers to remotely access their approved project data in a secure and auditable environment.

5.4 The study team's exclusive use of the data

Where a researcher (member of the ALSPAC team or an external collaborator) has secured funding for the collection and analysis of new data, they are entitled to apply for a period of exclusive access for a period of up to 6 months from the point at which a cleaned dataset is made available to them. If this is approved then during the exclusive access period the ALSPAC Executive will still consider requests for access to the restricted data, but permission must be sought from the researcher who funded the data collection to release the data or to explore the potential for collaborative analysis. If the funding researcher declines, the restricted data will not be available to others until after the period of exclusive access. After the embargo period all ALSPAC data are freely available to external researchers (within the usual constraints related to scientific legitimacy and disclosure risk). ALSPAC do not "police" overlap of projects or data requests. Details of approved projects and their data can be viewed by researchers on the ALSPAC website.

5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

The ALSPAC policy on data sharing is partly determined by the terms of the consent given by the participants to the collection of particular data items. Broadly, we work with consent agreements that allow the widest possible sharing of ALSPAC information within the scientific community, balanced against the need to recognise participant concerns that may influence their decisions about giving or withholding consent at the time of data collection. In the majority of cases, the anonymity of participants is maintained by providing linked data that do not include actual or potential personal identifiers (such as date of birth, name and address) and by minimizing potentially disclosive information (such as low cell counts). The point at which data become sufficiently detailed to the extent that anonymity cannot be preserved is sometimes unclear and subject to challenge by different parties. Where such situations arise, ALSPAC will take appropriate action to identify any risk to participant anonymity and where necessary take steps to alter the data or introduce additional stages to the research process to reduce these risks.

Depending on the proportion of the cohort consenting to different types of data collection, it is likely that ALSPAC will hold particular sets of data that were collected without specific individual consent. However, subsequent use of that data must preserve individual anonymity and thus not introduce any risk of inadvertent disclosure. To ensure this, ALSPAC may modify the data provided to a researcher in order to control for potential disclosure. ALSPAC are actively engaged with technological data sharing solutions that allow analysis across linked datasets but do not allow the analyst to have sight of either linking identifiers or the detail of individual level information needed for deductive disclosure, e.g. UKSeRP.

5.6 Regulation of responsibilities of users

The full ALSPAC data access policy is available online and provides information on data sharing for prospective researchers. In brief, researchers wishing to use the ALSPAC resource complete an online proposal form (https://proposals.epi.bristol.ac.uk/) describing the proposed research. The proposal should have clearly stated aims and hypotheses and describe the relevant exposure, outcome and confounders that are being requested. A Principal Investigator with an approved project is required to sign a Data Access Agreement (signed at an Institution level) and all researchers within that project must complete a confidentiality agreement before data is released. This emphasizes the confidential nature of the data and informs the researcher that they must not share their dataset nor attempt to match their dataset with any other ALSPAC data.

Requests to access biological samples are handled using the same procedures. However, the majority of samples represent a finite resource, so proposals are assessed to ensure analysis will make good use of these samples. Attempts are made to combine analyses where possible, but we reserve the right to turn down proposals which would use up a large proportion of finite samples. Samples are issued under the terms of a material transfer agreement (for researchers outside the University of Bristol) or a material service level agreement (for UoB researchers outside the Bristol Medical School (PHS)). Samples and indeed any other data are provided on condition that all further data obtained from them is returned to ALSPAC to become part of the ALSPAC resource and made available to other researchers. Where requests may use up a finite resource or risk stock, the request is referred to the ALSPAC Independent Scientific Advisory Board.

6. Responsibilities

The ALSPAC Executive Committee takes ultimate responsibility for all aspects of data management. Under the 2019-2024 strategic award the data team will be headed by the Executive lead for Data. They will be

supported by a Senior Data Manager (SDM) who manages the data pipeline and a Technical Lead (TL) who manages the ALSPAC systems. The SDM and TL will line manage a data team who will be responsible for managing ALSPAC systems that support the collection, curation and storage of data in a systematic, secure, confidential and accessible manner. The Executive Lead has responsibility for the final sign-off of new data released for research use and its' accompanying metadata. The SDM will be supported by a small team of data preparation assistants who will prepare and document the research data and a team of research associates who will provide continual quality assurance of new clinic data as it is collected. Dedicated time will be provided for data security; ensuring that we continue to comply and gain re-certification of ISO27001.

7. Relevant institutional, departmental or study policies on data sharing and data security

Policy	**URL or Reference**
Data Management Policy & Procedures	University policy: http://www.bristol.ac.uk/research/environment/governance/research-data-policy/
Data Security Policy	ISO27001 (reference)
Data Sharing Policy	ALSPAC access policy available here: http://www.bristol.ac.uk/alspac/researchers/access/
Institutional Information Policy	http://www.bristol.ac.uk/infosec/policies/
Other:	
Other	

8. Author of this Data Management Plan (Name) and, if different to that of the Principal Investigator, their **telephone & email contact details**

The ALSPAC executive

Email: alspac-exec@bristol.ac.uk Tel: +44 (117) 331 0010