1. Sources of data a. Questionnaire data files

Each questionnaire in the ALSPAC study in represented by one individual data file. There are six types of questionnaire. The files are generally referred to by the one/two letter stem shown in capitals:

- Mother (A, B, C…)
- Child-based, completed by the mother (KA, KB, KC …KX; TA, TB…)
- Partner- (PA, PB, PC…), Father (FA)
- Child completed. (CCA, CCB, CCC…)
- Teacher completed (SA, SB, SC…)
- Puberty (PUB1, PUB2, PUB3…..)

Each question response from a questionnaire is given a variable name, which is always prefixed with the letter stem of the built file name. So, for example, all variable names in the KB file start with KB (KB115, KB592a etc.).

b. Focus visit data

Children in Focus (a 10% sub-sample examined in more detail between 4 and 61 months of age) data is held in the CIF built file, which follows exactly the same format as those described above. The variable names are all prefixed with CF. As with the questionnaire data, complete documentation is available.

For Focus (child and parents) visits there is an individual built file for each time point, containing all of the visit information and each session's data. Please see the relevant built file documentation for detailed information on the variable naming.

c. **The cohort profile and MZ/KZ Files**

These are very important files that do not fall into the general data structure described above. Work is currently ongoing to combine the cohort profile and the KZ file into one. The files contain baseline information on all mothers and children who are part of the *eligible* cohort, such as birthweight, gestation and maternal age. The MZ file contains mother-based information and KZ child based. Complete documentation is available describing the information held on the KZ and MZ files.

The latest version of the cohort profile documentation (v2b) is available on Wellcome Open Research: https://wellcomeopenresearch.org/articles/4-51/v1.

d. Biological samples and other data

As well as the questionnaire and Focus data files, data are available that has been collected from other sources, including biological samples, delivery data and sub studies. Full documentation is not always available in Acrobat form for these data

sources, but is being prepared. Information regarding what data is available can be obtained from your data buddy if it is not clear.

2. How ALSPAC data files are linked internally

Cases in mother-based and partner-based files are identified by a unique five-digit number called ALN (ALSPAC Linking Number). ALN is also present in the child-based files, along with a second identifier, called QLET, which is required to identify children from multiple births. QLET takes the value A for singletons and the first born in a multiple birth, B for the second born in a multiple birth. Therefore each singleton has its own unique ALN but a pair of twins will have the same ALN but distinct values of QLET.

Note that QLET cannot be used to identify multiple births, because a value of A indicates singletons as well as the first born from a multiple birth. To identify multiple births the variable MZ010 is needed.

ALN and QLET facilitate the linking of all ALSPAC data files. For reasons of confidentiality, any research data which has this identifier cannot be linked back to the administrative database containing any contact details such as names and addresses.

External collaborators will **NOT** receive ALN, instead they will receive an identifier called a collaborator ID (CID), which is unique to each individual project. Individual children will therefore be identified using the relevant CID variable and QLET.

3. How the data are stored

The data are stored internally in SPSS system files and STATA data files. Each row in a data file represents one case (i.e. one child or one mother) and each column represents a different variable or questionnaire number.

Table 1: Form the data takes in SPSS

		Variable name				
ALN	**QLET**	**KA001**	**KA002**	**KA003**	**KA004**	**KA005**
30001	**A**					
30002	**A**					
30003	**A**					
30004	**A**					
30004	**B**					
30005	**A**					

Table 1 illustrates the layout of a child-based file. The mother represented by ALN 30004 has twins, so the responses to questions A001 to A005 may therefore be different for each of those children. In a mother-based file, there would be only one row for ALN 30004, since data on the mother would be the same for twin A and twin

B. *It is essential that QLET is used in the matching process when child data are involved* so that the same information for the mother is matched to the data for both her twins.

4. Data Documentation

All built files are fully documented. For all data, the source is described in detail (e.g. for questionnaire data the full text of the question is given, for Focus visits full details of each test are provided), with the corresponding variable names and frequencies of the responses. Details of derived variables are also provided in the documentation where available.

5. Data security

Access to ALPSAC data will only be provided after completion of a Data Access Agreement (DAA) and users will be bound by the terms of that agreement. The following points are important to note:

a. Storing ALSPAC data

ALSPAC data must be held on secure computer systems which are regularly backed up and restricted to authorised users only. ALSPAC data must never be stored locally on desktop or laptop computers. **b. Data transportation**

All data transferred electronically must be encrypted using AES-256 encryption (this can be achieved using compression tools such as WinZip or 7-Zip). This applies to all methods of electronic transfer, e.g. email may be used to transfer information but only when the classified data is an attached file encrypted to AES-256 or above. When using removable media, only USB flash drives with hardware AES-256 encryption must be used. Removable media should be transported securely and not be left unattended at any time.

c. Passwords for Encrypted Data

When sharing encrypted data, the encryption password must be exchanged by a separate mechanism than that used to transfer the data itself (i.e. do not send it to the same email address from the same email address). Encryption passwords must contain:

- a minimum of 10 characters;
- a combination of uppercase and lowercase letters;
- numbers;
- special characters (~!#@+=%).

Encryption passwords must not contain any single word found in a dictionary in any language.

d. ALSPAC Data and Information Security

If at any time the information security of ALSPAC data is breached e.g. data is lost, an unauthorised user gains access, etc; your ALSPAC Data Buddy must be notified immediately.

6. Sample Size and the new cases a. Core ALSPAC sample

The **eligible** sample comprises 20,248 pregnancies resulting in 20,390 known foetuses. The **enrolled** *core* ALSPAC sample consists of **14,541** pregnancies. This is the key number that should be quoted in all publications and represents Phase I enrolment. As described in the documentation for the MZ & KZ files, this is the number of pregnancies for which the mother enrolled in the ALSPAC study and had either returned at least one questionnaire or attended a "Children in Focus" clinic by 19/07/99.

The 14,541 core pregnancies resulted in **14,676** fetuses. 195 pregnancies were twin pregnancies, 3 were triplet pregnancies and 1 was a quadruplet pregnancy. Note that of these 14,676 fetuses, 14,062 were live births and 13,988 were alive at 1 year.

b. New cases

When the oldest children were approximately 7 years of age, an attempt was made to bolster the initial sample with eligible cases that failed to join the study originally. A further recruitment exercise from the age of 20 onwards resulted in further participants joining the study

As a result, when considering variables recorded from the age of 7 onwards there are data available for more than the 14,541 pregnancies mentioned above.

The total sample size for analyses using child-based data collected after the age of 7 is therefore **15,658**. Of this total sample of 15,658 fetuses, 14,975 were live births and 14,901 were alive at 1 year.

c. Sample sizes for collaborators

The dataset you receive from your data buddy will contain **15,645** cases, regardless of time points included. – for confidentiality reasons, data on the 13 triplet / quad children are not allowed to leave ALSPAC so will not be included in your dataset. Indicator variables will be included representing the core cases and those recruited during Phases II and III. If your dataset is only based on mothers, you will receive the 14,541 core pregnancies.

Other sample sizes may be encountered if other data sources (e.g. schools, DNA) are used as data may be available for eligible case with whom ALSPAC has had no direct contact. These will be detailed in separate documentation.