JISC/MRC Data Management Planning: Synthesis Report

Catherine Jones & Juan Bicarregui, STFC and Peter Singleton, Centre for Health Informatics and Multi-professional Education (CHIME), University College London

November 2011

1. Introduction

The JISC funded MRC Data Management Planning (DMP) project has examined data management planning issues in MRC-funded projects. The Science and Technology Facilities Council (STFC) and University College London, who were part of the consortium undertaking the Data Support Service (DSS) project for the MRC, carried out the work based with the DSS case studies with whom they had already built relationships. The MRC DSS project developed templates and guidance in Data Management and guidance on Data Sharing as part of its remit. The JISC/MRC DMP project asked the case studies to trial these templates and give feedback in order to refine them. This report describes that work.

The Case Studies involved were the Avon Longitudinal Study of Parents and Children at the University of Bristol; the Whitehall II Study at University College London and the MRC Unit in The Gambia.

2. Background

Although data management planning is applicable to all part of the MRC funding programme, the JISC/MRC DMP project focussed on Population Health studies.

2.1 Discipline description

Epidemiology or population cohort studies have several unique properties when considering data management planning. To be able to realise their scientific worth, information needs to be gathered from the cohort being studied over a long period (often several decades) and analysis done over data collected at sweeps of data collection, often separated by many years. This has several significant consequences related to data management[1]:

- data need to be understandable in the long term
- data need to be understandable by study team members who may not have been involved in the collection.
- the study/project needs to be aware of changes in standards and best practices for data management as their data may be a mix of paper-based and electronic data collection methods.

- [1] the data collected can include biological samples such as blood and more complex medical information such as MRI scans. However, we here focus on data provided as text.

- for continuity and comparison across different data collection waves, attention needs to be paid to the questions asked and protocols used, so that variables collected in different waves have the same meaning.
- studies that spread over many decades and waves of collecting data will have thousands of variables to manage, along with the coding books, questionnaires and other material required to understand and use the data in the future.
- all studies need to have ethical approval to undertake data collection.
  Cohort members will have signed consent forms which cover the collection of data and these may have an effect on what and how data may be shared in the future. For long-lived studies then it may be difficult, or impossible, to contact cohort members to re-consent under new terms and conditions which would enable uses that were not anticipated when the study was started. Due to the individual nature of the data, there are extra responsibilities on the principal investigator (PI) and their team regarding the data protection and confidentiality of the data. When sharing, either data strict procedures to control for the risk of participant disclosure need to be applied, or the requesting study needs to be vetted to ascertain that any consent from the data subjects would cover the requested study, considering the appropriate level of data security and other privacy controls..

2.2 Data Management Planning challenges

Funding for large scale population studies can be for a fixed period with scientific objectives set at the start, or for a particular data collection. When applying for funding, it is then a difficult task to distinguish between the data management required for new data collection against the requirements to ensure that data already collected remains valid and usable.

2.3 Data Sharing challenges

Within the Population Health Sciences there is a general acknowledgment that data sharing is beneficial and results in greater collaboration between academics in the field. However, there is some caution regarding the way in which data sharing application can be implemented and a general preference that data must be held within the studies rather than in a central repository, which would make the sharing activities harder to track and measure.

As with any research done using people, the Study leader is protective of the study participants, the way in which the study is perceived by those participants and is therefore mindful of the potential for research to run counter to an individual's beliefs or reputational damage. This is understandable as to continue to collect data; the study needs to retain the trust and support of the participants.

Ensuring that the data which is requested to be shared does not inadvertently identity individuals means that there needs to be carefully considered measures in place to control for disclosure risk.

2.4 Activities undertaken in the project

The project held a workshop in May 2011 where representatives from the Population Health community supported by the MRC came together to discuss the proposed Data Sharing Guidelines and Data Management Planning template and guidelines produced by the DSS partners University

College London and the STFC. Following a very productive discussion, and building on the work done in producing the DSP and DMP, three of the project present agreed to work with the project to produce sample data management plan templates to test the template, assessment process and the guidelines produced. These are reproduced in the Appendices of this document.

This synthesis considers the results of this process, together with suggested improvements to the form.

3. Data Management Planning Context

3.1 MRC plans for the evaluation of Data Management Plans

Data management planning covers the entire life-cycle of the project and the data it generates; the DMP project concentrates on the information required by the MRC as a funder at the point at which a proposal is evaluated within the MRC. The information gathered will be evaluated by a panel whose expertise is in the science rather than data management and thus the form needs to be succinct and give an opportunity to compare proposals. The science and resourcing involved in the proposal are covered in other forms required in the process. Following best practice a successful project would then produce a more detailed data management plan as the project continued.

3.2 Other funders' requirements and best practice

The sections below are a synopsis of the policies of funders in similar disciplines together with an overview of the domain-neutral guidelines provided by the Digital Curation Centre. As is to be expected there is a lot of communality over the key requirements for effective data management and sharing planning at the proposal stage.

3.2.1 USA's National Science Foundation (NSF)

http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

- Separate document of less than 2 pages of content
- Explain conformance to NSF policy on dissemination and sharing of research results
- Specific additional advice for certain domain areas.
- Topics to be covered may include:
    - Types of data, samples, physical collections, software and other materials to be produced
    - Standards for data & metadata content where in existence or project approach
    - Policies for access and sharing, including provision for appropriate protections of privacy, confidentiality, security, intellectual property or other rights
    - Policies & provisions for re-use, re-distribution or production of derivatives
    - Plans for archiving research products and preservation access to them

From the specific guidance from Social, Behavioural and Economic division:

PIs should use the opportunity of the DMP to give thought to matters such as:

- The types of data that their project might generate and eventually share with others, and under what conditions
- How data are to be managed and maintained until they are shared with others
- Factors that might impinge on their ability to manage data, e.g. legal and ethical restrictions on access to non-aggregated data
- The lowest level of aggregated data that PIs might share with others in the scientific community, given that community's norms on data
- The mechanism for sharing data and/or making them accessible to others
- Other types of information that should be maintained and shared regarding data, e.g. the way it was generated, analytical and procedural information, and the metadata

NSF FAQ's http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp

3.2.2 USA's National Institute of Health (NIH)

http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

The requirement for a Data sharing plan is limited to those who are Investigators seeking $500,000 or more in direct costs in any year

- The precise content of the data-sharing plan will vary, depending on the data being collected and how the investigator is planning to share the data. Plan may include
    - Brief description of the expected schedule for data sharing,
    - the format of the final dataset,
    - the documentation to be provided,
    - whether or not any analytic tools also will be provided,
    - whether or not a data-sharing agreement will be required and, if so, a brief description of such an agreement (including the criteria for deciding who can receive the data and whether or not any conditions will be placed on their use),
    - and the mode of data sharing

3.2.3 BBSRC – statement on data sharing

http://www.bbsrc.ac.uk/web/FILES/Policies/data-sharing-policy.pdf

- Separate document of one side of A4
- Have a statement on data sharing
- Conformance/compliance with BBSRC's Data Sharing Policy or explicit reasons why this is not possible
- Concise plans for data management
- Plans may include details of:
    - Data areas and data types (volume, type & content)
    - Standards & metadata
    - Relationship to data available in public repositories
    - Secondary use

- Methods for data sharing
- Proprietary data (restrictions)
- Timeframes for data release
- Format of the final dataset

3.2.4 Cancer Research UK -data management and sharing plan

http://science.cancerresearchuk.org/funding/terms-conditions-and-policies/policy-data-sharing/data-sharing-guidelines/

- Required to produce a data sharing and management plan for a grant – adhering to policy.
- Areas to consider are:
  - Volume, type, content & format of the final dataset
  - Standards used for data collection and management
  - Metadata and documentation for correct interpretation
  - Method for sharing data
  - Timescale for public release
  - Long-term preservation plan
  - Whether a data sharing agreement will be required
  - Restrictions on potential sharing: exploitation, proprietary data, confidentially, ethics or content

3.2.5 Wellcome Trust - data management and sharing plan

http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/index.htm

- Research proposals should include approach to management of data generated
- Plan required where proposal involves generated of data that have clear scope for wider reuse and long term value.
- Plan should be clear and concise and avoid repetition of content from elsewhere in the proposal.
- Contents should address the following questions:
  - What data outputs will your research generate and what data will have value to other researchers?
    - what types of data the proposed research will generate
    - which data will have value to other research users and could be shared
    - what data formats and quality standards will be applied to enable the data to be shared effectively.
  - When will you share the data?
  - Where will you make the data available?
  - How will other researchers be able to access the data?
  - Are any limits to data sharing required - for example, to either safeguard research participants or to gain appropriate intellectual property protection?
  - How will you ensure that key datasets are preserved to ensure their long-term value?
  - What resources will you require to deliver your plan?
    - People and skills - is there sufficient expertise and resource in the research team to manage, preserve and share the data effectively? Is additional

specialist expertise (or training for existing staff) required? If so, how will this be sourced?

- Infrastructure - are there appropriate computational facilities to manage, store and analyse the data generated by the research?
- Tools - will additional computational facilities and resources need to be accessed, and what will be the costs associated with this?

3.2.6 ESRC– data management plan

http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx

All ESRC grants applicants who generate data need to subject a data management and sharing plan

- Explain the existing data sources to be used and gap analysis between existing data and proposed data collection.
- information on the data that will be produced by the research project, including the following:
  - data volume
  - data type, e.g. qualitative or quantitative data
  - data quality, formats, standards documentation and metadata
  - methodologies for data collection
- planned quality assurance and back-up procedures [security/storage];
- plans for management and archiving of collected data;
- expected difficulties in data sharing, along with and causes and possible measures to overcome these difficulties;
- explicit mention of consent, confidentiality, anonymisation and other ethical considerations;
- copyright and intellectual property ownership of the data; and
- responsibilities for data management and curation within research teams at all participating institutions.

3.2.7 DCC Data management plan checklist headings

Version: v2.2: 6th January 2010

1. Introduction & context

The first section of the DMP checklist is explaining the project and policy context. For a DMP in a grant proposal, most of this content would not be required specifically as it would already be in the paperwork. Recommendation is that the **Name of the project** and **document creation date** are mandatory.

1. Legal and ethical issues

2. Access, sharing & re-use

Policy questions around sharing the data, how the data will be made available and restrictions on reuse.

3. Data standards and capture methods

Description of the data/data types; where the data is coming from, file formats & metadata

4. Short term storage and data management

Volumes, back-ups etc.

5. Deposit and long term archiving

Addressing the requirement for long-term archiving, persistent citation and preservation.

6. Resourcing

7. **Adherence, review and long-term management**
8. **Agreement/ratification by stakeholders (if useful)**
9. **Annexes**

4. The first draft Data Management Template

A pilot DMP template for use in applying for MRC funding was produced (see below) based on best practice within the sector and considering the generic tools provided by the Digital Curation Centre.

The template is intended to be completed to produce a DMP which sits alongside the proposal which covers the scientific case for funding and its associated resourcing. This DMP is designed to consider data management and sharing issues alone. The aim of the DMP is to inform the evaluators without overwhelming them with detailed information.

The template is divided into four main sections:

- Data description and policy links
- Data acquisition/collection
- Data management, curation and preservation
- Data sharing

These are preceded by a section which deals with existing policies which may be at institutional, department or group level. It is expected that the panel would be able to see these policies if required, but they need not necessarily be open – for example the data security policy is unlikely to be publicly available.

The pilot template is shown below.

Proposal name/number	
URL link to the policies referred to in this proposal Please add any others required	
Policy	**URL**
Data Sharing Policy	
Institutional Data Management Policy	
Project/Study Data Management Policy & procedures	
Data security policy	
Data Description overview Describe the key aspects of the data relating to this funding proposal in no more than three lines of text.	

Data Acquisition and/or collection This section should address the following areas: a) What is the type, format and scale of data to be collected? b) What standards will be used for the data and associated metadata? Highlight areas of innovation from standard practice in your field
Management, Curation and Preservation This section should address the following areas: a) What retention period is proposed for these data? b) What strategies & standards will be used for managing and curating the data and associated metadata?
Collaboration and Sharing Describe the key principles & practices from your Data Sharing Policy: Specifically: • The method for sharing data • Proposed timescales for public release • Standards for providing the data • Whether a data sharing agreement will be required • Restrictions on potential sharing

Date/timescale for release of data	
Data sharing agreement required?	YES/NO
Further information:	
Reference paper/index paper available from:	

The accompanying guidelines expanded on the sections in more detail. The expectation is that the review panel would be able to access the policies during the review process, not that the policies are necessarily publically available.

5. The Revised Data Management Template

5.1 Case Studies

Three MRC funded studies agreed to trial the MRC data management plan form. These were: Avon Longitudinal Study of Parents and Children (ALSPAC); the Whitehall II Study and MRC The Gambia. ALSPAC and Whitehall II are both longitudinal studies which started in 1991 and 1985 respectively. MRC The Gambia chose an example of a clinical trial.

5.2 Feedback on the template

The template was discussed at the workshop on MRC data management planning held in May 2011; whilst there was general agreement about the concept there were comments on how it might be improved.

Comments and suggestions for improvements:

- Difficult to describe the data collected longitudinal study in three sentences to any level of detail.
- Should the policy also include direct reference to the consent policy/practice?
- *Data* could be seen as narrow term, those completing the form may think that materials such as video are outside of the form's remit.
- Checklists would help both those who have to fill it in and those who might have to evaluate them. Particular areas that might benefit from checklists were: types of data collected, standards to be used (especially security ones)
- When does data collection stop and data cleaning start?
- There was a lot of discussion about tools and techniques for locating potential data and also how these might be kept up to date by studies and the different audiences.

One Data Manager suggested some changes to the form to make it more useful and for it to align with the information she feels is necessary. See section 7.2 for her example using her modified format. The main changes were to add a checklist of potential types of data which can be collected and a new section on data security measures. This reflects both the type of data one can collect about individuals and the importance of ensuring data about them is managed in a secure and confidential manner.

5.3 Discussion on the proposed amendments

This section discusses the amendment of the template section by section both from the point of view of cohort studies and then from a wider MRC perspective.

1) Introduction

Proposal name/number
Data Overview
Describe the key aspects of the data relating to this funding proposal in no more than three lines of text (resource, number of subjects, number of collection waves, %participation, etc).

The start of the revised form brings the introductory material together and puts the policy information to the bottom of the form. This makes for a more logical sequence for completion.

2) Data Collection

Data Collection

1) Data types

Qualitative	☐	Genotypic data	☐
Quantitative	☐	External Mortality/Medical records	☐
Interview/Home visits	☐	External Administrative records	☐
Clinical measurements	☐	Tissue samples: Blood, DNA, Urine	☐
Self-completion questionnaires	☐	Images (ECG, MRI, etc)	☐
Personal/confidential data	☐	Other _____	☐

2) Format and scale of the data (no of data tables, no of records, or total file size in Mb, Gb, Tb, or Pb)

3) Methodologies for data collection and data quality control. Please specify standards used, if any.

The original *data acquisition and/or collection* is renamed "data collection" and the response is more structured by way of the use of a checklist.

This would make for easier input and comparison. However the data types are aimed at cohort studies and the aim of the template is to cover all types of MRC funded research. The feedback during the template design phase was that checklists make it easier for the PI to fill in and to meet the aims of the form.

Possible ways forward could be to:

1. *Add to the data types check list to include all types of data covered by MRC science.*

   This would require effort to ensure all standard types of data are included and might mean that the form was full of options which were not applicable to the PI. On the other hand it would mean that everyone would complete the same form.

2. *Have the supplementary guidance have subject-specific checklists which can be added to the electronic form by the PI.*

In this case there would be specific domain checklists and it would be the responsibility of the PI to add it to the form. The benefit would be that there would be one form, but this would put more responsibility and effort onto the PI.

3. H*ave separate forms for different types of science.*

   The benefit would be that each form would contain relevant information in the checklist. The downsides are maintaining more than one form and the problems of cross-discplinary science.

4. *Have a separate form for the long-term cohort studies and for other grant proposal applications.*

   This would leave the existing suggested template as it is for all grant proposals but provide a specialised form for long term cohorts in recognition that data management planning for these studies is more involved and complicated.

   This concept could be extended to have an optional separate form for clinical studies where the interest may be on the phase of study, and the data modules included such as Demography, Adverse Events, Concomitant Medications, PK/PD etc.

There were the following comments on the precise details of the check boxes:

- Qualitative: I wonder whether the category should be 'mainly qualitative' and 'mainly quantitative' as most of our small surveys usually have both qualitative and quantitative
- The layout makes these appear as distinct (non-overlapping) 'data types' whereas they are more 'data attributes' which may apply to some or all of the data.
- There is a high probability that all the data is 'personal data' as unlikely to be assuredly anonymised – and so a better definition of what was meant here would be appreciated.
- Are tissue samples 'data' – there may be data about the tissue samples, which is different

3) Data Curation and Documentation

Data Curation and Documentation
a) <u>Strategies for managing and curating data. Please specify standards used, if any.</u>
b) <u>Strategies for documenting data. Please specify metadata standards used, if any.</u>
c) <u>Strategy for data preservation. What retention period is proposed for these data?</u>

The original management, curation and preservation section is renamed data curation and documentation but the subject matter covered is the same.

There is a philosophical question as to when data should be documented, by placing the questions in a section entitled "curation and documentation" it could be inferred that documentation is an activity which is not part of the collection phase. However for the most effective curation and re-use some documentation or recording of metadata should be done at the collection phase.

Feedback from one of the case studies was "My experience: Documentation is done throughout the collection phase for things like protocols, questionnaires, forms, etc. However, documentation of the resulting datasets (metadata, data dictionaries, syntax used, cleaning procedures, etc) is done way after the data collection has finished, when data curation has been done, derived variables are in place and data manager has the time to focus on documentation. It might be better to have a section for "Data Curation and Data Preservation" and another specifically dedicated to "Data Documentation", in order to avoid any philosophical issues?"

4) Data Security and Privacy Controls

Data Security and Privacy Controls
Describe the key aspects of the strategies followed to ensure data security and preservation of confidentiality (security measures, access rights, preservation of confidentiality, etc).

This is additional section not covered in the original template. It would be interesting to discover whether this particular aspect is of interest to the whole MRC funded community and as such should form part of a standard template. The feedback received from our case studies show that it is welcomed. In particular as it encourages scientists to consider carefully the implication of 'restraints' such as the Data Protection Act and the need for Anonymisation etc together with any associated costs.

There was a suggestion that this section should be made more specific by asking the following questions:

a. Which of the following approaches to data-sharing is used:

i. Restricted record-level data (possibly 'personal data') to approved studies
ii. De-identified record-level data subject to data-sharing agreement
iii. Record-level data may be linked in-house or at trusted third-party (TTP)
iv. Researcher may have access at host institution, subject to honorary contract and supervision
v. Study may submit queries and aggregate or derived results returned
vi. Study must work through host researcher

vii. Other – please detail:

b. Various security controls:

i. How is access controlled? (e.g. eGIF)
ii. Are there facilities for remote access? What additional restrictions apply to remote access?
iii. Is personal data/identifying detail separated from clinical data? How is access to this restricted?
iv. Are separate 'local' backups made and these tracked via a secure log?
v. Are there audit trails on access?
vi. How are audit trails monitored to identify possible misuse or intrusion?
vii. Are these all covered in your Data Security Policy and procedures?

5) Data Sharing

Collaboration and Data Sharing
Describe the key principles & practices from your Data Sharing Policy:
1. Method for sharing data:
Full open access ☐ Informal exchange ☐ Gated access ☐ Restricted access (on-site) ☐
2. Data sharing agreement required? YES ☐ NO ☐
3. Proposed timescale/date for public release
4. Procedure followed for providing the data
5. Restrictions on potential sharing

The additional of the check boxes for data sharing reflect the domain from which the suggestions have come and may not be directly transferrable to other parts of the MRC funded community.

The MRC's policy on data sharing is "The MRC expects valuable data arising from MRC-funded research to be made available to the scientific community with as few restrictions as possible so as to maximize the value of the data for research and for eventual patient and public benefit. Such data must be shared in a timely and responsible manner."[2]

[2] http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/Policy/index.htm

The suggested changes in this section generated the most discussion within the case studies. The suggestion of tick boxes without any further description wasn't unanimously agreed with. Concerns were expressed about what exactly "full open access" meant and whether it would be interpreted in the same way by all PIs and whether by having the tick boxes rather than an implicit assumption that data will be shared unless there are good reason might encourage conservative scientists to opt for restricting access more tightly.

If a checkbox approach is to be taken, then it needs to reflect all mechanisms used by the communities funded by MRC.

URL link to the policies referred to in this proposal (where publicly available online) Please add any others required	
Policy	**URL**
Data Management Policy & Procedures	
Data Security Policy	
Data Sharing Policy	
Institutional Information Policy	
Other:	
Other	
Reference paper/index paper available from	

This section has been moved from the start to the end of the form. It makes sense to have this at the end.

5.4 Comments on the completed templates

The ALSPAC and Whitehall examples show that for a long-running and complex project it is not possible to compress the amount of data management information into the suggested two pages guideline. MRC The Gambia, with the clinical trial, expressed the required information within the suggested limit, leading us to suggest that for most MRC funded proposals 2-3 pages will be sufficient. Whitehall have used the revised data management plan template.

5.5 Revised Template

Considering the scope, the workshop comments, previous feedback, the revised form is below:

Proposal name/number
Data Overview
Describe the key aspects of the data relating to this funding proposal in no more than three lines of text (resource, number of subjects, number of collection waves, %participation, etc).

Data Collection

1. Data types

Some kind of checklist – see discussion in 5.3

☐ ☐

2. Format and scale of the data

3. Methodologies for data collection and data quality control. Please specify standards used, if any.

4. Information on the metadata to be created and the process for generating it

Data Management, Documentation and Curation

1. Strategies for managing and curating data. Please specify standards used, if any.

2. Strategies for documenting data. Please specify metadata standards used, if any.

3. Strategy for data preservation. What retention period is proposed for these data?

Data Security & Privacy Controls

Describe the key aspects of the strategies followed to ensure data security and preservation of confidentiality (security measures, access rights, preservation of confidentiality, etc).

This could be made more specific by adding the two lists described in section 5.3 to this form, or encouraging this by adding the lists into the guidance.

Collaboration and Data Sharing

Describe the key principles & practices from your Data Sharing Policy:

1. Describe the method and procedures for sharing data:

Will a Data sharing agreement be required? YES ☐ NO ☐

2. Proposed timescale/date for public release

3. Data excluded from sharing and justification (i.e. ownership, consent

4. Restrictions on access to data

URL link to the policies referred to in this proposal Please add any others required	
Policy	**URL**
Data Management Policy & Procedures	
Data Security Policy	
Data Sharing Policy	
Institutional Information Policy	
Other:	
Other	
Reference paper/index paper available from	

6. Conclusions

The aim of this work is to ensure that data generated from public funds is well managed so as to maximise the value of the science made possible by it. Requiring those who wish to be funded to make data management plans at proposal stage ensures that it is considered from the start.

- Using a template means that it is easier for the PI to complete and for the panel to evaluate rather than the free-form methods adopted by other funders in similar domains.
- However this may be not be the most appropriate method of assessing the data management strengths of cohort studies who are funded over a considerable length of time
- Even a short synopsis of the information about the management of the data runs to more than the two sides of A4 which was the original brief for the template.

The project team had feedback both with the template design phase and the completion phase about the utility of check boxes, both for those who had to fill in the template and to assist the evaluation of it.

- We noted this but it is difficult to provide forms with lots of checkboxes for a funder such as the Medical Research Council which has such a wide variety of domains and standards within them.
- Section 5.3 *Discussion on the proposed amendments* suggests four ways of addressing this usability feature.

By using the case studies for this project as initial completers of the template it has enabled the project team to both adjust for minor changes and make recommendations to the MRC about improvements to the template. As with all forms, the initial version needs to be tested under the wide variety of potential uses before it settles into the final form which gives benefit to both sides.

The MRC intends to take this work forward this autumn and to align this with new policy in this area.

7. Appendices

7.1 ALSPAC form

FOR EXEMPLAR PURPOSES ONLY: To demonstrate the use of this template with a large programme support grant we have used many elements of the ALSPAC 2011-2015 Wellcome Trust/MRC core programme support application. However to demonstrate other aspects of this form we have added hypothetical information that does not relate to the actual grant. Therefore this document is for exemplar use only.

Proposal name/number	
The Avon Longitudinal Study of Parents and Children (ALSPAC): An international resource for population genomics and lifecourse epidemiology. Core Programme Support 2011-2015 Submitted via Wellcome Trust. Trust reference: 046734 eGrants reference: 15298	
URL link to the policies referred to in this proposal Please add any others required	
Policy	**URL**
Data Sharing Policy	http://www.bristol.ac.uk/alspac/sci-com/collab-policy/
Institutional Data Management Policy	
Project/Study Data Management Policy & procedures	Currently Under Development (Due Summer 2011)
Data security policy (DSP)	
- Institutional DSP	http://www.bristol.ac.uk/infosec/policies/
- Departmental DSP	http://www.epi.bris.ac.uk/intranet/policy/policy.htm
- ALSPAC DSP	Currently Under Development (Due Summer 2011)
Data Description overview Describe the key aspects of the data relating to this funding proposal in no more than three lines of text.	
ALSPAC is a two-generation, sample of ~14,500 mothers ('M') recruited in pregnancy (1990-92), their offspring (O) and partners (P). The resource; a biobank, genome-wide DNA (M, O n=10k, P n=1k), clinical assessments (M n=1, O n=9), questionnaires (M n=30, O n=40, P n=10), linkage to records.	
Data Acquisition and/or collection This section should address the following areas: a) What is the type, format and scale of data to be collected? b) What standards will be used for the data and associated metadata? Highlight areas of innovation from standard practice in your field	
This proposal will further develop the ALSPAC resource in ways that render it of maximal value to the scientific community. The resource will be developed by (1) A questionnaire to collect data on the mother and partner as they move towards later middle age. i. Type: Participant completed questionnaire with an e-questionnaire option. ii. Format: Electronically captured (scanned paper or digitally collected) data with each data item assessed by logical and range checks, ambiguous values assessed by an operator. All assessment scales to have been validated externally with a known reference paper. iii. Scale: Four questionnaires, each ~20 pages in length. O n~10,000, M n~10,000, P n~4,000. (2) Recruitment grandchildren (G) of the mothers (M) originally recruited to create a unique three-generation resource. (3) Recruit siblings (S) of the original ALSPAC children to enter the study and piloting low-cost data collection from them. i. Type: Recruitment campaign with follow up consent and data collection interview. ii. Format: Electronically captured (scanned or digitally collected) enrolment and consent (for record linkage) forms. High quality administrative data, including where possible administrative IDs such as NHS ID number. Interview data to be collected and validated in real time on encrypted laptops with data transferred to the central repository routinely.	

iii. Scale: G 1000 cases (estimate) will be born in the grant period. The parents indicating they wish to participate will need a consent and data collection interview and consent paperwork. So, it is difficult to assess the likely enrolment rate for this group.
(4) A clinical examination and further biological sample collection on the fathers (P);
i. Type: A clinical assessment visit hosted at the University of Bristol, ALSPAC premises.
ii. Format: Personal visit, each participant's progress through sessions identified by a unique ID and recorded using custom administrative software. System data and image files will be transferred routinely to the central data server. Sample data to be transferred routinely to the ALSPAC LIMS (laboratory information management system). Data collected will include biological samples, interview data, computer collected questionnaire data, scan image files, scan data files.
iii. Scale: P 3,500 cases (estimate). The clinic will take 1.5 to 2 hours.

Metadata will be collected as an integral process to (i) catalogue and index the data in a searchable manner, (ii) define the assessment tools (scales, key reference publication, modifications etc), (iii) to describe the data collection process on an individual basis (age at completion, administration and reminder process) and to (iv) assign a geographical reference point (at a non-disclosive level) to assist spatial analysis. The metadata standard (i.e. DDI) that ALSPAC will adopt has not been finalised.
Data will be collected in line with the departmental data security policy. All data collected will be catalogued and archived in the ALSPAC SAN (storage area network), which is scalable, secure and backed-up routinely.

Management, Curation and Preservation
This section should address the following areas:
a) What retention period is proposed for these data?
b) What strategies & standards will be used for managing and curating the data and associated metadata?

The ALSPAC Data Management Policy is defined elsewhere (see URL) and is described here in brief. Data from all sources have been cleaned, validated and prepared for analysis by the in house statistics and data management team. Each data item is referenced and stored using a universal indexing and naming convention. Data items are stored separately from administrative data (including subject identifiers) and linked through anonymised linkage files accessible only to specified members of the data management team. All data sets are maintained on secure servers with a regular backup policy. A full audit trail is maintained of all changes made to the data and is available as part of the documentation. Our security policy is described in the included URL. We are currently undertaking a data security GAP analysis and a series of internal and external audits that will underpin a further security upgrade to achieve ISO 27001 certification before the end of the next strategic award period.
ALSPAC is envisaged to operate indefinitely and to operate as a resource for current and future generations, in line with this expectation all data are anticipated to be maintained indefinitely. Where data are disposed of (for example data, that has been secured elsewhere, on an obsolete hard disk) this will be done securely and in line with our IT Information security policies.
ALSPAC is fully engaged with the MRC Data Support Service project, running until April 2011, which will lead to the provision of study metadata in a fully searchable form. ALSPAC has already provided the MRC DSS with extensive descriptive metadata of available information, and our intention is to ensure that all our published resource metadata are accessible at this location. We are also committed to work with other national initiatives to provide direct access to our data. In particular, the ESRC funded Secure Data Service has been identified as a widely used conduit for making data available to bona fide researchers. We are committed to using centrally managed systems to deliver these services rather than developing bespoke ALSPAC systems.
The ALSPAC web site will be used to host data documentation and catalogues.
The University of Bristol have arranged to indefinitely secure and host the ALSPAC paperwork as one of the University libraries 'Special Collections'.

Collaboration and Sharing
Describe the key principles & practices from your Data Sharing Policy:
Specifically:
The method for sharing data
Proposed timescales for public release
Standards for providing the data
Whether a data sharing agreement will be required

Restrictions on potential sharing	
Date/timescale for release of data	1 year from the end of data collection
Data sharing agreement required?	Yes

Further information:

Currently, ALSPAC data is made available to collaborating scientists on the basis of a supported rather than an unrestricted, open resource. Bespoke datasets of requested variables are provided to collaborators by a data preparation and statistics team. Our current policy on data sharing and access is described on the web at the URL listed above. Briefly, available data are described on our website. Researchers wishing access to these data currently submit a proposal to ALSPAC Executive, which considers these with a target turnaround time of 2 weeks. Approval is given if the proposal is scientifically sound and feasible (ALSPAC has the required data). Specified variables are then provided to the investigator by a 'data buddy' (ALPSAC statistical team), who supports the user with data descriptors and additional variables as required. Data are provided as SPSS, STATA or CSV files. Projects that generate new data, including de novo measurements and variables derived from existing data, are expected to deposit all variables in the central information reservoir along with appropriate documentation to allow their incorporation into the ALSPAC data dictionary.

Where a researcher has secured funding for the collection and analysis of data then they are offered a period of exclusive access, lasting for up to 1 year from the point at which the data are made available. During this period the ALSPAC Executive will consider requests for access to restricted data and seek approval from the researcher who funded the data to release the data or to explore the potential for collaborative analysis. In the case of data requests for projects that overlap with current work, we encourage collaboration between the relevant parties, however this is not a barrier to accessing data.

Requests for genetic variables are processed in a similar manner and governed by a data transfer agreement between the University of Bristol and the collaborator's host institution (http://www.bris.ac.uk/alspac/documents/appendix5-dta-2008-08.pdf). Genomewide association analyses are currently conducted in Bristol and summary data are prepared for collaborators on the basis of agreed protocols.

In future, within the lifetime of the strategic award under review, ALSPAC is committed to providing open access to ALSPAC data to the widest possible research community.

Our policy on data sharing is partly determined by the terms of consent 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 many cases of data use, the anonymity of participants is maintained by providing linked data that do not include actual or potential personal identifiers. The point at which data become sufficiently detailed to the extent that anonymity cannot be preserved is sometimes unclear, but is likely to be influenced by processes, such as linkage, that enrich the database.

Intellectual property rights belong to the University of Bristol. We will consider dividing intellectual property rights where researchers outside of the University of Bristol will be making a particular contribution. Any such division must be considered and agreed before the research starts. Further information on the University of Bristol policy on intellectual property can be found on: http://www.bris.ac.uk/research/knowledgetransfer/ip/ipownership.html.

Depending on the proportion of the cohort consenting to different types of data collection, it is likely that ALSPAC will come to hold a body of data collected without individual consent that can consequently only be used in scientific applications that preserve individual anonymity and do not introduce any risk of inadvertent disclosure. All users of ALSPAC data will be required to work within these constraints.

Reference paper/index paper available from:
http://www.bristol.ac.uk/alspac/documents/appendix1-methodology.pdf

8. Whitehall form

Proposal name/number
Whitehall II

Data Overview

Describe the key aspects of the data relating to this funding proposal in no more than three lines of text (resource, number of subjects, number of collection waves, %participation, etc).

Whitehall II is a closed occupational cohort. At recruitment in 1985 it included 10,308 participants, all employees of the British Civil service; 3,413 women and 6,895 men. To date the cohort has been subject to 9 data collection waves. At wave 9 the participation rate was 66% of wave 1 responders.

Data Collection

1) Data types collected (tick all that apply)

Qualitative	☐	Genotypic data	☑
Quantitative	☑	External Mortality/Medical records	☑
Interview/Home visits	☑	External Administrative records	☐
Clinical measurements	☑	Tissue samples: Blood, DNA, Urine	☑
Self-completion questionnaires	☑	Images (ECG, MRI, etc)	☑
Personal/confidential data	☑	Other _____	☐

2) Format and scale of the data

Paper copies:

- Personal details: correspondence, contact details forms, consent forms, etc
- Completed questionnaires in the UCL Records Office (off-site) or in the Whitehall premises.
- Electrocardiograms taken at phases 1, 3, 5, 7 and 9
- External data: GP notes and hospital notes from tracing of CVD and stroke medical events.

Electronic databases:

- Access databases with personal information of participants and details of the participation.
- Completed questionnaires: microfilms for phase 1, PDF copies for phases 2 to 9.
- SAS/Stata database for statistical analysis: questionnaire data, clinical results, genetic data and derived variables.
- External data: Hospital Episodes Statistics, mortality data, cancer registration (supplied by the NHS Medical Research Information Service), myocardial infarction data from MINAP.

Biological samples:

- Blood samples (from waves 3, 5, 7 and 9), DNA samples (from waves 7 and 9) and urine samples (from phase 1) are kept in UCL freezers.

3) Methodologies for data collection and data quality control. Please specify standards used, if any.

- Piloting of questionnaire and clinical data collection.
- Weekly monitoring of clinical information received. Self-completion questionnaire data checked every 6 months.
- 10% of participants have all measurements repeated to check reliability of methods used
- 5% of biological samples are split and blinded to check the data quality received from the laboratories responsible

for analysing the samples

- Double entry of clinical data, screening forms, questionnaire responses and participant consent.

Data Curation and Documentation

a) Strategies for managing and curating data. Please specify standards used, if any.

- Participants identified with the anonymised ID. Consistent across different collection waves.
- Data quality checked by monitoring inconsistencies, outliers and missing data.
- Access databases are used for clinical and administrative data.
- The core research datasets are curated in SAS. The organisation of these datasets has been done using a normalised relational database design approach. Labels and formats as descriptive as possible.
- Syntax of derived variables and naming of variables consistent across collection waves.
- Established file management strategy (folder structure and file naming).

b) Strategies for documenting data. Please specify metadata standards used, if any.

- Excel data dictionaries across waves. Variables are categorised for easy discovery.
- Copies of questionnaires and screening forms in PDF format
- Documentation of the syntax of derived variables
- Protocols and coding manuals in RTF and PDF formats
- Metadata from the actual SAS and Stata datasets, through labels and formats.
- XML based online data dictionary

Documentation is done from the start of the collection wave.

Additionally, SAS programs have been developed to extract metadata in XML format with the ultimate aim of representing Whitehall metadata using the DDI standard (Data Documentation Initiative).

c) Strategy for data preservation. What retention period is proposed for these data?

- Please see security measures.
- Retention period: indefinite.

Data Security

Describe the key aspects of the strategies followed to ensure data security and preservation of confidentiality (security measures, access rights, preservation of confidentiality).

Security measures

- Secure storage of data in the Whitehall premises. Secure transfer of files and paper work.
- Secure and regular backups. Disaster recovery arrangements rely on the backup policies.
- System Level Security Policy approved by the NIGB.

Internal controlled access

Within the Whitehall premises, different access rights have been granted to Whitehall staff depending on their role within the group. Off-site access by Whitehall researchers is enabled through the centrally managed UCL network. Only specific research datasets can be stored in this network.

Preservation of confidentiality

Whitehall data are classified into three categories:

1) Confidential data (number, name, contact details, DOB, etc.) are exclusively handled by the Administrative team

2) Linked anonymised data: only used for internal research. Anonymisation is achieved using the study ID. Confidential variables such as names are removed, but fields of sensitive nature are still present (detailed ethnicity, job title, dates of medical events, etc), which could potentially lead to a disclosure of individual identities.

3) 'Full' anonymisation (low level of disclosure risk): provided to external researchers as part of the data sharing policy. Extra restrictions been put in place to protect the potential disclosure individual identities: i) double anonymisation of subjects; ii) a different ID set for each approved project; iii) de-identification of data by removing

personal information and variables with low prevalence rates.

Collaboration and Data Sharing

Describe the key principles & practices from your Data Sharing Policy:

1. **Method for sharing data:**

Full open access ☐ Informal exchange ☐ Gated/controlled access ☑ Restricted access (on-site) ☐

2. **Data sharing agreement required (yes/no)?**

Yes

3. **Proposed timescale/date for public release**

All data from the previous collection wave are publicly available as soon as the data collected at the current collection wave are curated and ready to be analysed. E.g. wave 8 data were made available as soon as wave 9 data were curated.

4. **Procedure followed for providing the data**

We have adopted the controlled access approach to prevent misuse of data through unethical, premature, opportunistic or poorly-designed analysis. Data are only available to *bona-fide* researchers with projects of high scientific probity. External researchers are required to submit a formal application outlining their research plan to and listing the data needed. Upon approval of this application by the Whitehall Scientific Committee, applicants are asked to sign a Data Sharing Agreement indicating their compliance with Whitehall II data sharing terms and conditions. Following the receipt of this document, the data manager releases an anonymised dataset, tailor-made for the specific project.

A separate data sharing policy is in place for genetic projects, which involve either genotyping of Whitehall DNA samples or request existing genetic data.

The Whitehall II controlled access model for data sharing conforms to WHO/Wellcome Trust code of conduct of sharing public health data (www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology).

5. **Restrictions on potential sharing**

We ask the applicant to re-submit their application if:

- Unclear or vague description of the project
- Request of data we do not have
- Project with very large scope that largely exceeds the recommended1-2 papers within the first 2 years
- List of variables containing information that is not specified in the project description

URL link to the policies referred to in this proposal

Please add any others required

Policy	URL
Data Management Policy & Procedures	Not available online. PDF file available on request.
Data Security Policy	Not available online. PDF file available on request.
Data Sharing Policy	www.ucl.ac.uk/whitehallII/data-sharing
Institutional Information Policy	http://www.ucl.ac.uk/isd/common/cst/swg/policy
Other:	
Other:	

Reference paper/index paper available from

??

9. MRC The Gambia form

Proposal name/number	
AN OPEN RANDOMIZED PHASE I STUDY EVALUATING SAFETY AND IMMUNOGENICITY OF A CANDIDATE HIV-1 VACCINE, MVA.HIVA, ADMINISTERED TO HEALTHY INFANTS BORN TO HIV-1/2-UNINFECTED MOTHERS	
URL link to the policies referred to in this proposal Please add any others required	
Policy	**URL**
Data Sharing Policy	http://(draft)
Institutional Data Management Policy	http://(draft)
Project/Study Data Management Policy & procedures	http:// (draft)
Data security policy	http://(draft)
Data Description overview Describe the key aspects of the data relating to this funding proposal in no more than three lines of text.	
Clinical and immunological data from 48 Gambian HIV uninfected mothers and infants vaccinated at 20 weeks old. Study not powered to provide definitive evidence of vaccine-induced effects; rather it is based on smaller numbers of participants and will only indicate trends, feasibility, or size of effect and variability of measurements to inform the design of more definitive studies. The case report form has 34 pages of data which is double entered into a study database, Immunogenicity data is managed separately within Excel spreadsheets.	
Data Acquisition and/or collection This section should address the following areas: c) What is the type, format and scale of data to be collected? d) What standards will be used for the data and associated metadata? Highlight areas of innovation from standard practice in your field	
a. Data includes Laboratory (Haemotology, Biochemistry, MVA.HIVA Immunogenicity, EPI Immunogenicity, HIV test and HLA typing), informed consent, physical examination, vital signs, inclusion and exclusion criteria, sample collection, adverse events, serious adverse events, concomitant medications collected at 4 clinically planned visits. The case report form has 34 pages and the study will run from October 2009 until June 2011. b. Project data standards have been developed with Kenya study site (sister study but with HIV positive Mothers). Data was collected into OpenClinica which is an Open Source electronic data capture and data management system, this system is metadata driven in that the CRFs and database are both built from a Microsoft Excel spreadsheet input which contains comprehensive metadata. Innovation was to use OpenClinca at the MRC the Gambia as this was the first study we used the system for.	

Management, Curation and Preservation

This section should address the following areas:

c) What retention period is proposed for these data?

d) What strategies & standards will be used for managing and curating the data and associated metadata?

a. Data will be archived for 15 years. The paper records will be held in the MRC Unit the Gambia Archive (air conditioned and with appropriate pest control processes) and the electronic data will be held in a SQL Server database on the Archive server called 'Condor'.

b. The Archive has SOPs in place detailing the strategy for keeping the records safe.

Collaboration and Sharing

Describe the key principles & practices from your Data Sharing Policy:

Specifically:

- The method for sharing data
- Proposed timescales for public release
- Standards for providing the data
- Whether a data sharing agreement will be required
- Restrictions on potential sharing

Date/timescale for release of data	December 2014
Data sharing agreement required?	YES

Further information:

Emerging data will be shared within the study teams in Kenya and Gambia. The summarised results will also be entered on ClinicalTrials.gov database following completion of the trials.

Safety and immunogenicity data from this vaccine study will be shared more widely within 3 years after completion of all assays; assays are still ongoing.

Reference paper/index paper available from: