EUCCONET Scientific Report

Data Management Interest Group – Workshop 9-10th March 2010

Summary

Cohort studies are complex projects that generate vast amounts of data over long periods of time. These data need to be organised and secured in a way which enables the long-term follow up of the study families, archiving of data and the provision of data for complex research purposes. The challenges faced by data teams in cohort studies across Europe are similar, however these staff are potentially less likely to meet and interact with their peers as they are often administrative or contract based rather than researchers.

The call for establishing a 'Data Management Interest group' came from a meeting of the EUCCONET steering group in Paris in February 2009. This first meeting of the steering group debated which topic areas from the discipline of cohort studies would benefit from the establishment of a peer networking group and the use of EUCCONET resources to promote the group, and group aims, with a workshop.

The 'Data Management Interest Group' group aims are:

- Establish a group and provide a forum for specialist data staff from child cohort studies to discuss topical issues and to identify and share best practice.
- To use the information gained to encourage the development of templates outlining 'generic' systems of use in cohort studies and the identification and development of standards within the field.

The 2010 workshop provided a starting forum for data specialists to meet and discuss their methodologies, the barriers they face and solutions they have developed. The workshop was aimed primarily at data managers as well as specifically database administrators, database designers and data processors from the bio-informatics field. 42 delegates representing 12 European child cohort studies attended. To broaden the knowledge and experience present at the workshop the field was expanded to include two US panel studies as well as representatives from 'NatCen' and 'ScotCen', UK based social research centres.

Speakers and topics were selected from nominations made from the delegates. Sessions comprised of an introductory presentation outlining systems or topic areas, followed by round table discussion. To address the aim of developing a ‘generic system template’ it was decided to identify and focus on a system that is common to almost all cohort studies– the 'administrative' database. This system is used to maintain a record of contact and participation of cohort members and generally underpins the mechanisms used to distribute data collection tools. System design was debated at a parallel sessions aimed at the database designer and supported by a survey sent to all studies in advance. The session concentrated on identifying common elements of a 'generic' system and tried to identify best practice.

Feedback from delegates suggests that much was learnt from the workshop. It also suggests that many delegates felt that they had few previous opportunities to meet with their peers and to debate these issues at a technical and specialist level. In particular the discussion around designing a generic administrative database proved to be popular and suggests a model with which to discuss other areas in the future. Additionally topics including 'open data access' and the future requirements of bio-informatics were felt very useful and sparked interesting debates and exchange of experience and proposed solutions.

The workshop outputs meet the aim of establishing a working group of data specialists. Group contact and discussion will be encouraged through group emails and the establishment of a dedicated discussion forum, via the EUCCONET website. The information gathered from the 'administrative database design' sessions will be collated into a report. This report will be made available as a reference document from the EUCCONET website. Further promotion of the group and dissemination of findings will be made via application to conferences for presentation opportunities. The EUCCONET and methodology session of the inaugural Society for Longitudinal and Life Course Studies conference has been identified as an appropriate venue.

EUCCONET Scientific Report

Data Management Interest Group – Workshop 9-10th March 2010

Scientific Content and Discussion

Design

The 2010 Data Management Interest Group (EDMIG) was designed to provide a starting forum for data team staff to meet and discuss the various methodologies utilised and identify best practice from the data teams across, primarily, European child cohort studies.

A key feature of the workshop design, and the EUCCONET programme, was to seek presentations from the delegates. Communication with group members sought to set the agenda of the meeting, however due to the enormous breadth of 'data management' it was not possible to cover all aspects of the discipline that might have been of interest to the delegates. The programme settled upon a range of topics, looking in depth at 'Data Access', 'Database Design' and Bio-informatics as well as a range of single topic presentations.

Delegates

The initial interest group design was to attract specialist data team staff. These study personnel were envisaged to work mainly with data and need not fulfil any research function. Identified 'key' personnel to attract to the group included:

- 'Data Managers' - staff with an overall responsibility within a study for the collection, processing, documentation and archival of data.
- 'Database Administrators' - staff with a responsibility for designing and maintaining the database systems that generally underpin the administration of a cohort and are increasingly being used for data storage and distribution.
- 'Bio-informaticians' - staff with a responsibility for processing, cleaning and storing large scale, primarily genetic, datasets.

In practice it was quickly noted that each study had a distinct approach to managing these data needs. This approach was largely determined by study size and resources but was also strongly influenced by the length of time the study had been running. Outsourcing of these functions to contractors was also a common feature, particularly in specialist areas such as database design, programming and security. The outcome of these varying approaches was that delegates from large, well established cohorts could be categorised into the intended target group however delegates from smaller studies or studies in the early stages of design or implementation, tended to be those members of staff who were assigned these roles or would have the role of implementing data elements of the study design.

42 delegates attended the workshop. Delegates represented 12 European child cohort studies; two US panel studies as well as two UK social research centres that facilitate data collection and processing at UK cohort studies.

Presentations & Follow on discussions

The workshop was opened with an introductory session outlining the aims of the group, the workshop and brief introductions from delegates. A delegate from each study centre made a brief, informal, introduction to their study outlining the size of the cohort, the primary data collection mechanisms, if bio-samples were being collected and the current study status (i.e. design, recruiting, active data collection etc.).

Primary Topic Area 1: Data Access

Jon Johnson (CLS - UK): Birth Cohort and Panel Study Data Management and Documentation in the USA and England: An overview of the findings from the Survey Resources Network and pointers for future directions. Review of range of approaches to data management – from loosely to highly integrated systems.
This presentation raised issue of trying to find standards that can be used by all cohorts. The need for a data management equivalent to Dublin Core for longitudinal/birth cohort studies.

Paul Snell (ALSPAC - UK): Moving towards open access data
Inger Meder (Danish National Birth Cohort - Denmark): The Danish National Birth Cohort – Data resources, linkage and data access

Discussion following these presentations centred on the pros (good end user support, controllable), and cons (resource intensive, can cause a bottleneck) of ALSPACs supported access system and moves to identify an open access system. Debated systems included the Danish CITRIX example, Nesstar and UK existing solutions including 'Secure Data Services'. Concerns were raised about the quality and integration of derived variables. There was some review of the 'Inquisite' on-line system of questionnaires used by DNBC to gather data.

Primary Topic Area 2: Administrative Database Design

Administrative databases serve a core function in cohort studies of all designs. These databases are the repositories of the personal details of cohort members, record the participation status, vital status and to enable the mechanics of distributing data collection tools to the cohort members. The database design is critical as flexibility and expandability is paramount in enabling staff to administer dynamic study designs and complex family situations, including divorce, fathering children across multiple family units, the study young people having children of their own.

The workshop sought to explore this area using a mix of a short survey sent out in advance and discussion sessions held in parallel to the main workshop targeted at the database specialists from each study. The sessions followed the structure of the survey materials to debate this topic.

A discussion point that highlights the importance of ensuring that these systems are well thought out and flexible is that several studies have had to redesign their database, and associated applications, to enable changing data collection requirements and evolving family dynamics.

The survey results and discussion will be used to compile a report on this subject as one of the primary outputs of the workshop. However the results can be summarised as:

- Almost all studies used the commercial relational databases provided by Oracle or Microsoft. The hardware and software choice is not seen as critical provided a well supported relational database is used.
- These databases support applications written in a variety of software languages. These applications are either deployed as stand alone applications or are web based
- Many studies maintained full time staff that specialised in databases and application development. The use of specialist contractors was common but no study used an 'off the shelf' commercial package.
- Most studies made a distinction between 'administrative' and 'research' data and did not store research (i.e. questionnaire) data collected from cohort members within these systems.
- The ID numbers used to identify each cohort member were either generated in house or studies made use of nationally implemented ID numbers such as social security IDs. ID numbers were either allocated at an individual basis or at a pregnancy level accompanied with birth order codes and family relationship codes. Delegates provided some steer towards allocating IDs at an individual level as this may provide additional flexibility. There were strong recommendations that ID numbers include a 'check-sum' algorithm to allow systems to check for data entry errors.
- The majority of studies use national standards and look-up tables to ensure the validity of cohort addresses. Delegates discussed the advantages in linkage, GIS and administrative efficiencies gained by using these standard formats. This was seen as an area worthy of further investigation as studies noted the difficulties in retrospectively establishing stable geographical indicators.
- Many studies routinely collect nationally implemented Ids, such as health service ID numbers, social security numbers. This is seen as being of great benefit in aiding efficient and accurate data linkage.
- Security is seen as a key function of the database system. Security is enforced through a mixture of the general IT infrastructure, the features contained with the database package and mechanisms enforced by application design and staff working practice.

Primary Topic Area 3: Bio-informatics

This session was led by researchers with an understanding and interest in data issues. It was decided to invite guest speakers from outside of the delegate list as the intention was to look to upcoming trends over the next 5 years.

Dr Nic Timpson (MRC CaiTE – UK): Future directions in genetic epidemiology, impact on IT and Data requirements

This presentation was followed by considerable discussion primarily about storing and accessing these data. It was recognised that as genetic testing becomes cheaper and faster to turn around then cohort studies will be routinely conducting GWAS and potentially whole genome scans. This and the developments in areas such as expression data will place considerable strain on existing systems. Discussion identified that the time scales for this were short to medium term and that there would be considerable pressure placed on data teams due to the competitive nature of genetic research h and the desire to enter into consortia. While some studies had

limited experience with GWAS data none had experience of these large data sets across the whole cohort. Group consensus suggested that delivering these large data repositories was achievable within a cohort study but would need to be factored into the infrastructure planning cycle. These systems would need to be flexible and easy to expand; one study was actively sourcing a SAN system to meet these requirements.

Dr Sue Ring (ALSPAC – UK): Changing practice and regulations and the impact on laboratory IT requirements

Dominic Hoff (MoBa – Norway): The MoBa LIMS application

These two presentations looked at the changing demands of laboratory IT requirements and how these were being driven by legislative and ethical requirements. These required the existence of extensive audit trails that extend from sample location and identification to the participants consent form. The MoBA example tightly integrated the LIMS system with the administrative cohort database. Studies had found few obvious off-the-shelf solution; MoBa's was developed in house, another study had commissioned a system in conjunction with other departments from their host University and another was investigating a specialist package.

Assorted Topics:

i. **GIS**

Andy Boyd (ALSPAC - UK): Developing a GIS resource within a cohort study

Following this presentation there was a discussion concerning the difficulties that had been found across various studies in retrospectively assigning stable geographical indicators. There was debate about both the appropriate scale at which to do this (neighbourhood or household) and the additional resources and complexity inherent in recording location at a household level.

ii. **Security**

Anthony Philips (Security Consultant – UK) : Meeting international security standards within a cohort study

Unfortunately the scheduled presentation was cancelled at a late stage due to ill-health. Many delegates stressed the importance of this topic and that it merited discussion at a later date.

Plenary & Feedback:

Delegates expressed a desire to document consensus, or even a set of 'standard's, on data management best practice. This was the intention behind the 'Administrative Database Design' topic which was received well. Studies in the design phase highlighted the importance of information of this nature. It was noted that legislative differences between the study countries may hinder establishing consensus in some topic areas, such as open access.

Delegate feedback suggests the workshop met its goal of providing an opportunity for data managers to meet and discuss common issues at a technical level. Feedback confirmed the thinking that 'data managers' as a group may have fewer opportunities to meet and share best practice than their research colleagues:

Delegates expressed a desire to meet again. It was decided that another group member should take this forward and apply to EUCCONET for funding for a follow up workshop. However no individual has been nominated to date.

EUCCONET Scientific Report

Data Management Interest Group – Workshop 9-10th March 2010

Results & Impact on the Field

Assessment of Results:

1. Network

The workshop was the first major initiative of the Data Management Interest Group and aided the promotion of the group's interests and aims. In early EUCCONET distribution materials only a handful expressed an interest in joining the group. Following the workshop publicity a total of 42 delegates attended, representing 12 European child cohort studies. To broaden the knowledge and experience present at the workshop the field was expanded to include two US panel studies as well as representatives from 'NatCen' and 'ScotCen', UK based social research centres.

The workshop was run in an informal manner which was designed to encourage interaction. The use of unstructured time in the schedule encouraged much discussion and the building of a group dynamic. This assessment of these workshop aims is best left to the feedback:

"I felt there was a genuine sense of collaboration between delegates – with the desire to learn from others and give advice to others"

"for a group that has not met before, there were very productive discussions"

"I have been involved in collecting data in public health science for seven years. This is the first time I have met so many people responsible for the data handling in such projects discussing the complexity of the task. I think that bringing these people together was in itself a fantastic idea"

Delegate lists and contact details were distributed to all attendees.

2. Forum

While feedback suggests that the workshop was successful in building a network of data specialists this must be maintained. It is proposed that this is enabled through group emails and the use of the EUCCONET on-line forum. This should be promoted regularly. To help establish this forum as a working resource the organisers will start discussion based around the future direction on the field and on establishing a 'generic' template for an administrative database that reflects the best practice identified by the group.

"Being a data manager for large studies is very challenging and those using your data – the PhD's running the studies often cannot help provide guidance in the practicalities of data manipulation and storage issues – Learning from others doing the same work was very worthwhile."

3. Developing a 'generic' administrative database design.

This session was received very well with delegates. Participation in the survey was high (10 out of 12 studies returned the survey) and there was a lively discussion in the sub-group session, While there was inevitable variation in system design amongst the studies it was possible to identify common elements. Where variation occurred there was constructive discussion to identify the strengths and weaknesses of each design. The group agreed that it seemed possible to establish a clear set of advice and best practice guidelines that would form the basis of a generic system template.

"I especially liked the exchange between experts on Database Design. Normally you do not get this opportunity at other conferences."

The report is still under preparation and will be discussed amongst group members before it is finalised. It is envisaged that the report will be made available on the EUCCONET website and the group will consider if it is appropriate to disseminate the findings through publication or presentation at a suitable venue.

4. Future direction of the group.

Conclusions reached in the plenary discussion and in individual feedback suggest that group members

found the opportunity to debate these topics at a specialist and technical level was of great benefit. The group thought it would be of benefit to apply for funding to hold another workshop in the future. It was suggested that a member of the group should come forward to lead this application. Although no member was identified at the time this will be followed up via email and on the forum. It was agreed that EUCCONET was an appropriate source of funding and that an application should be made to the steering committee.

Impact on the future direction of the field:

1. **Network**

The workshop, and in future the interest group, provides a network for data specialists to share experience to identify solutions and best practice. It was noted that many data issues are of common interest across child cohort studies and where best practice is put to use efficiencies can be made. The hindsight and experience of established studies is of great use to new studies in the design phase and conversely the use of new technologies available to new studies could have benefits across the field.

> "As we are at a very early stage of our study this gave me great insight into the methodological challenges that we will face and how other groups have dealt with these"

> "It was great to see so many data managers together. ...I think it is a very good start for further collaboration between the child studies in Europe."

2. **Forum**

The EUCCONET on-line forum represents an easy to access and low cost resource for group members. This however relies on the interaction by members and their good will. Feedback suggests that the workshop has achieved a good precedent for this; however the forum will need considerable promotion for it to achieve its potential.

3. **Best Practice & Standards**

While this is perhaps the most ambitious of the group aims it also has the potential to have the greatest impact on the field. The experience of administrative database design exercise illustrates that it is possible to fulfil this aim at a functional, advisory level. It would take significantly greater effort, potentially a dedicated workshop and follow up collaborative work to develop true standards that can be applied across the field. However it can be seen from delegate comments that there is a desire to investigate this:

> "Can we establish a 'Dublin Core' for Longitudinal / Birth surveys?"

> "It would be really great to get consensus from a group such as this on best practice for data management. ...It would be an incredibly useful basis for current and future studies"

> "If the group can take the lead in pushing for some 'standards' ... setting an ISO for birth cohort database handling and management?"

> The potential here ranges from reference documents (such as the generic database template) that can aid studies in the design phase to standards that can aid study convergence and aid the production of cross-cohort standardised data required by consortium based approaches.

4. **Wider Dissemination**

To maximise the potential benefits of this group the experience and outputs should be disseminated to the field. While it is clear the findings should be made available via the EUCCONET website it would increase the impact of this group if it could present at a suitable conference or identify materials suitable for publication.

RESEARCH NETWORKING PROGRAMME

EUCCONET Data Management Workshop

Workshop Schedule

Tuesday 9th March

10.00 – Registration & Welcome Coffee

10.30 – 12.00 - Workshop Outline & Introductions

- Chairs Welcome & Workshop Outline
- All study centres to introduce themselves and their study

12.00 – 13.00 - Session 1: Data Management & Documentation

- Jon Johnson: Birth Cohort and Panel Study Data Management and Documentation in the USA and England: An overview of the findings from the Survey Resources Network and pointers for future directions

13.00 – 14.00 - Lunch

14.00 – 15.30 - Session2: Either Data Access OR Administrative Database Design

Data Managers	Database Designers
Data Access	**Administrative Database Design**
• Paul Snell: Moving towards open access data • Inger Meder: The Danish National Birth Cohort – Data resources, linkage and data access	• All DBAs to present a short (5 – 10 min) overview of their administrative database • Discussion to identify common elements and establish best practice

15.30 – 16.00 - Coffee

16.00 - 17.00 - Session 3: Data Access

- Ingo Barkow: Controlling access to NEPS items via a user management system

19.30 - Evening Meal at the Bristol Lido

Wednesday 10th March

9.00 – Coffee

9.30 – 12.00 - Session 4: Future Directions in Bio-informatics

- Nic Timpson: Future directions in genetic epidemiology, impact on IT and Data requirements
- Sue Ring: Changing practice and regulations and the impact on laboratory IT requirements

11.00 – 11.20 - Coffee

- Dominic Hoff: The MoBa LIMS application

12.00 – 13.15 - Session 5: Security & G.I.S

- Anthony Philips : Meeting international security standards within a cohort study
- Andy Boyd: Developing a GIS resource within a cohort study

13.15 – 14.15 - Lunch

14.15 – 15.30 - Plenary: Closing Discussions

- Future Direction of the Data Management Interest Group
- Closing discussion & workshop evaluation

15.30 - Workshop Close