Trials Units Information Systems – Data Standards

Data and Information Systems Project

The DIMS Project Team
Trials Units IS – Data Standards

Contents
Introduction .................................................................................................................................. 1

General Issues .......................................................................................................................... 2

The Scope of Data Standards .................................................................................................. 2
Current Use of Standards in Clinical Trials ........................................................................... 3
So Why Bother? ......................................................................................................................... 4
The Importance of the Trials Units ......................................................................................... 6
Central Organisations and Tool Development ....................................................................... 6
The Wider Trials Landscape .................................................................................................... 7
Change Management for Standards ......................................................................................... 8

Recommendations 1: Clinical Databases ............................................................................... 9

Using the Same Subject Identifiers ....................................................................................... 9
Common Clinical Question Sets .............................................................................................. 10
Common Non-Clinical Data Items ............................................................................................ 10
Clinical Data – Standardising Definitions and Responses ....................................................... 11
The Need for a Managed Data Item Library and a Selection Tool ............................................. 15

Recommendations 2: Trial Description and Metadata ............................................................. 17

Trial Descriptions .................................................................................................................. 17
Schedule Descriptions ............................................................................................................ 19
Question Descriptions .............................................................................................................. 20

Recommendations 3: File Management .................................................................................. 21

Data Collection Formats ......................................................................................................... 21
Collaboration Formats ............................................................................................................. 23
Long Term Curation of Data .................................................................................................... 25

Recommendations 4: Trial Management and Administration ................................................ 26

Pharmacovigilance ................................................................................................................... 26
Contact Data ............................................................................................................................ 27
General Trial Administration ................................................................................................... 29

Time Frames ............................................................................................................................. 30

System Vendors and Time Frames .......................................................................................... 33
Actions for the first 3 months .................................................................................................. 34
Actions for the first 12 months ................................................................................................. 35

List of Interviewees ................................................................................................................. 38

References ................................................................................................................................. 39
**Introduction**

This report is one of the planned deliverables of the UKCRC / NIHR DIMS (Data and Information Management Systems) project. Its aim is to explore how data standards can be used within clinical trials units to support both the science and the management of trials, to make recommendations as to which particular standards should be followed, and to make specific proposals for actions that could be taken in the short term to promote the use of data standards.

The recommendations made are based upon a review of the major standards and the literature surrounding them, a consideration of current and future contextual issues (e.g. the development of relevant systems in the NHS) and extensive consultation with key staff and opinion leaders, mostly drawn from the clinical trials community and the NHS, particularly the NIHR, but including some representatives from the pharmaceutical industry (the full list of interviewees can be found at the end of this document).

Examining data standards is a key element of the DIMS project but it is closely inter-related with the other areas – e.g. developing and applying system standards, and a general review of the architecture and specification of clinical trial IT systems. This report should therefore be read in conjunction with the others produced by the DIMS project team.

The report has three main sections, reflecting the aims listed above:

a) **General Issues:** This discusses the types of standards that could be applied, the current use of data standards in UK academic trials units, the key benefits that data standards could bring, and some contextual and organisational issues that will need to be borne in mind if standards are to be introduced successfully.

b) **Recommendations:** The second section makes recommendations concerning the most appropriate strategies to pursue with respect to standards, in each of four areas. The pertinent features of the main standards are described as necessary, but please note there is no attempt to provide a systematic review of all the available standards*

   c) **Time Frames:** Broad time frames are proposed for implementing different types of standards. These lead on to considerations of what initial preparatory work would be required, and then what actions would be necessary within the first 12 months of any standards implementation programme.

The focus therefore moves from a relatively general consideration of data standards at the beginning to a much more detailed, specific, list of suggested actions at the end.

---

* An annotated glossary is available separately that provides brief descriptions of most of the major standards and organisations, as well as links to their web pages.
General Issues

The Scope of Data Standards

The generic term ‘Data Standards’ can be (and is) interpreted in a variety of ways. In the context of this report the term is used as inclusively as possible, so that...

a) Data refers to all the types of data collected, stored and processed by trials units, i.e. both the scientific data garnered during the trials themselves and the administrative data used in managing those trials, and

b) Data Standards means any and all systems for promoting commonality and consistency between trials and between trials units, in the ways that data is selected, defined, structured, formatted, transformed and transferred.

This broad definition allows many activities to be considered as potential areas for data standards, as illustrated by Figure 1 below.

Figure 1: Potential areas for Data Standards in Clinical Trials Units
The twelve areas identified in Figure 1 are not a comprehensive list of all the potential areas for standardisation, though they probably represent the topics most discussed in this context (the details of each are discussed in the Recommendation sections).

The figure is intended to emphasise the diversity of potential standards. It underlines the fact that there is no single, over-arching standards system, no complete ontology for clinical trials, that could meet all the potential needs for standards. Nor would it be worthwhile to try and develop, let alone implement, such a system. Instead we have a heterogeneous set of existing and emerging standards which over time need to be linked together, as and when such linkage makes sense.

Figure 1 also illustrates that despite this diversity it is possible to divide the application of data standards into four broad domains:

a) **Clinical Databases**: This includes decisions about what data is actually sought and when, the ways the resultant questions are defined and coded, possibly with reference to formal vocabularies, the way data collection is structured (e.g. pre-defined data items versus user generated lists) and the way responses are coded and analysed, e.g. using controlled vocabularies such as those of SNOMED, MedDRA or LOINC.

b) **Trial Descriptions**: Various aspects of a trial need to be described to fully characterise it: for instance its overall aims and type, its identifiers, sponsors, sub-protocols etc; its schedule of visits, milestones and data collection events; and the detailed descriptions of each data item. A variety of methods and formats are currently used for this trial description and metadata – but they could be standardised.

c) **File Management**: The ways in which data is packaged, formatted and transmitted between different stages of the trial process. Examples are the way data could be passed back to the database from the primary data collection instrument, the way data is transformed from its primary state to a more standardised structure for submission to regulatory authorities, and the ways in which data is stored long term.

d) **Trial Management**: A range of data sets are used in managing trials. Some are purely administrative, such as people and organisations and their contact details, milestones in the trial approval and funding processes, and the research governance and agreement data of centres. Other data is more clinical, such as SAE reports and the resulting data flows between trials units, investigators, and regulatory authorities. All of it, however, could potentially be standardised.

These domains are clearly inter-related but distinct enough to be a useful way of classifying data standards in broad terms. They are therefore used to structure the recommendation sections of the report.

**Current Use of Standards in Clinical Trials**

Two of the questions in the recent consultation document sent to UKCRN registered trials units (TUs) related to data standards. 28 of the 40 trials units (70%) returned responses and those 28 appeared broadly representative of the group as a whole.

The replies to the questions about standards supported the impression obtained previously from informal discussion – that most trials units are aware of some data standards, though
only a minority currently use them, and that most units would be willing to investigate standards further to see if they could be useful to them.

The first question dealt with current awareness of data standards. Seven (25%) of trials units claimed a good knowledge of data standards, whilst 3 (11%) said they currently had no or very little knowledge of them. The remaining 64% had some awareness of some standards but were making no claims to be experts or particularly engaged with the issue.

The second question was about current use of standards, and here the most common response concerned MedDRA. 7 TUs (25%) said that they currently use MedDRA for SAE and/or AE coding, with a further 4 saying they were looking into using the system or anticipated having to use it in the future. Only three TUs (11%) said they were using WHO Drug, and three were using CDISC, though exactly how was not specified. Only 2 units used all of MedDRA, CDISC and WHO Drug. (Two further TUs reported using standards but were not specific about which and how.) In general the use of standards was more apparent in larger, longer established units.

It appears that there is currently a broad but sometimes vague awareness of standards, and of coding systems in particular, but actual usage is confined to a minority of trials units (~30% now, rising to ~40% in the future). Use of non coding standards, e.g. CDISC, is much more limited, at about 10%.

On the positive side, there were no overtly hostile remarks about using standards in clinical trials, though admittedly a couple that indicated indifference. Several respondents reported they were investigating standards further and/or expected to be using them more in the future. One explicitly asked for ‘central guidance as this is a complex area’.

So Why Bother?

It is clear, therefore, that although few people disagree with the notion of using data standards in clinical trials their current use is limited. This may be partly because the lack (so far) of a central steer or co-ordinated approach has made it very difficult to know which standards to use, but partly, it must be conceded, it is because few people find the issue of standards urgent or compelling.

Nevertheless the use of data standards is seen as an important part of the DIMS review. It behooves this report to rehearse the reasons why. Below are five (overlapping) reasons for increasing the use of data standards in clinical trials. The first four apply to trials as currently practised. The fifth reason relates more to trials as they might be carried out in the future (though it is the one which, possibly, may become the most important reason of all).

a) To increase the scientific and medical value of clinical trial data… by making the data generated by trials easier (and quicker and cheaper) to locate, characterise, share, and compare, for instance by using standard metadata description systems, and/or common sets of core questions. This would not only make traditional collaboration and meta-analysis easier, it could also open the door to new questions and analyses that are very difficult to do now. It would allow clinicians to more easily identify whether a particular trial, and/or that trial’s conclusions, were relevant to an individual patient. It would also allow the long term curation of trial data to be put onto a more secure footing.
b) To make trials quicker and easier to set up...by (for instance) using off the shelf question sets from a common library, common controlled vocabularies for coding, and encouraging common data structures and design philosophies. This would not only allow the specification of clinical trial databases and systems to become more standardised, it would also make it much easier to develop common tools that can support trial design and maintenance, and help to ensure compliance with documentation and other regulatory requirements.

c) To make trial administration more consistent and efficient...and management and central monitoring easier, and allow the specification and construction of tools designed to support these processes more transparent and efficient. This may be, for instance, by using common datasets for referencing the principal actors and locations involved in trials, by agreeing common data models for trial management, and by establishing standards for consistent data flow in trials units, e.g. with regard to pharmacovigilance or research governance.

d) To allow trial IT systems to become more modular and flexible...because the design, use, development and/or purchase of clinical trial IT systems is becoming more expensive and, as the regulatory framework increases in complexity, more demanding. If IT systems in trials units can be gradually re-engineered as assemblies of linked modules, necessarily using defined schemas and standard interfaces for data interchange, we can gain considerable flexibility in future system development.

In particular it would free trial units from the risks associated with large proprietary systems, and from the practical and political difficulties associated with any single, centrally specified, solution. A modular approach to design would also allow innovation to be more easily integrated into systems, and make it much easier to link trials units IT with systems elsewhere, for instance a central NIHR contacts database or a remote randomisation facility in Europe.

e) To allow the import of data from non trial sources...in particular from the routine clinical data captured in primary and secondary services. Despite the ethical, legal, logistic and security issues that need to be resolved, it has been widely recognised that the advent of electronic health records opens up the possibility of capturing ‘ordinary’ clinical data and using it for research purposes. This could completely transform the way in which trials are set up and conducted, at least with UK residents.

Connecting for Health’s Research Capability Programme (RCP¹), itself backed by the Treasury’s Office for Strategic Co-ordination of Health Research (OSCHR²), is testament to the seriousness with which this idea is being pursued. Using clinical data directly, however, is dependent upon systems in trials units ‘talking the same language’ – the same codes, the same XML schemas etc. – as those in the source systems. Even though importing data in this way is unlikely to happen any time soon, the potential importance of this development is so great for clinical trials that it is imperative that the relevant standards are understood and incorporated by trials units.

Data standards do not, therefore, represent some bizarre bureaucratic reflex to enforce conformity on scientific data sets. Indeed, whilst standardisation might mean conformity at the trivial level of (say) question codes, at a more fundamental level it is about liberating data, allowing it to be more easily moved, mixed, collated and compared with other data. Standards can also liberate trials unit staff, including IT staff, from some repetitive tasks – particularly
those involved in trial design and administration. As the fourth paragraph makes clear data standards can liberate systems as well, allowing them to be broken up into functional units as a prelude to more flexible, targeted and responsive development. Finally, standards, in conjunction with developments in NHS IT, could even liberate trials units from having to collect their own data directly, allowing them to draw research data from that generated from normal practice.

Implementing standards will always be an investment of current effort for future gain. In the past it has been difficult to justify the investment because there has been no central body to take the lead and where necessary support the process. The advent of the NIHR, the activity of the UKCRN’s IS working group, and the development of the RCP all mean that this is no longer the case – for the first time we can discuss, establish and implement standards seriously. As the reasons above make clear, the returns from a co-ordinated investment in standards could be pervasive and substantial. Far from being a peripheral ‘add-on’, data standards should be a central element of future IT systems in clinical trials.

The Importance of the Trials Units

Although individual trials units will be aware of the general and scientific advantages of data standards, some may base their decisions to implement them not so much on the technical merits of the standards themselves, as on:

- a) Considerations of costs (usually in the short term) and benefits (usually in the medium and long term), not just in terms of time and money but also in terms of the potential impact on working practices, staff training etc.

- b) Acceptability in terms of perceived ownership and control, provenance and credibility, and the recognition given to previously invested time and effort in this area.

- c) Incentives and support provided by central bodies and funders, and the perception that this is a long term change actively supported by central agencies.

- d) Evolving regulatory requirements and constraints imposed by central bodies and funders.

The heads of the trials units and their key staff, including the IT leads, need to see the overall benefits of standardisation – to their unit – as outweighing the costs and investment involved. If standardisation is viewed as an additional bureaucratic overhead imposed from the centre, for instance to make data collection more efficient or in order to achieve some ‘grand plan’ that has no relationship with day to day activity, it will have only very limited success. The potential benefits of data standards are very real, but they will need to be explained and argued over many times, as part of an ongoing, inclusive consultation process.

Central Organisations and Tool Development

Increased use of data standards should help national organisations such as NIHR, UKCRN, MHRA and NHS RCP achieve their own goals, but those central organisations must in turn be ready to support the development of standards. This may include funding additional workload incurred during pilot phases of standards introduction, and / or funding additional posts to support or co-ordinate the process – in fact any funding that will help tip the trial unit’s cost-benefit equation towards implementation.

A particular need will be for tools. Trials units will be much more likely to use standardised coding schemes, for instance, if they are supplied with a tool that automatically inserts those
codes in a new trial during the design phase. Similarly they are much more likely to connect to
and consume a remote repository of contact information if they are supplied with a client
program that can do just that.

Developing such tools is not something that necessarily has to be initiated by some central
(e.g. NIHR) team, though that would be an option. In some cases such tools may already be
developed, or in development, in one or more units around the country. In these cases a
central team might have more of a role in bringing together the expertise in trials units, co-
ordinating the development and validation of tools (or even purchasing them) and preventing
wasteful re-invention.

Where the input of the centre would be vital is in underwriting the continued maintenance
and support of tools and other supporting systems. Anything developed and supported by a
single trials unit, or even jointly by a few, will be vulnerable to staff and other local changes.
Central co-ordination will therefore be vital to ensure ownership of source code, adequate
documentation and continuity of expertise.

A successful data standards initiative must therefore include a commitment by the centre to
provide ongoing support, co-ordination and maintenance for the systems and tools that, in
turn, support that initiative in the trials units. Once systems are embedded, technically and
culturally, such central support may well change in character, but for the first five years (at
least) trials units will need to be assured that this is a programme that is fully supported by the
centre, in practical terms like code and component maintenance, and that it therefore has a
long term future.

The Wider Trials Landscape

Most academic trials units will continue to have most of their patients and collaborating
personnel based in the NHS. We therefore need to remain aware of, and where necessary liaise
with, those NHS initiatives e.g. NIHR programmes, CFH’s RCP and development of the
SUS, the Janet / N3 gateway, the use of SNOMED CT, and the new Cancer Intelligence
Network (and similar schemes that may follow in other disease areas), that will have direct
relevance to information flows and developing data standards.

It must be remembered, however, that most trials units will also have a proportion of their
patients and collaborators drawn from outside the NHS, from centres overseas and,
occasionally, the private sector. In addition collaboration with the pharmaceutical industry is,
and doubtless will remain, extremely important to many trials units.

Any standards that are developed should, therefore, work in the broader, global trials
landscape and not be confined to NHS compatibility, as important as that is. We should be
willing to examine and sometimes adopt approaches and systems used in the pharmaceutical
industry, partly because industry has already travelled further down the standards road than
most academic units, partly because that way trials units can remain attractive collaborators
for the drug companies, and can continue to enjoy the financial benefits such collaboration
often brings.

Standards and associated information systems are developing apace in both the
pharmaceutical world and the NHS. The two sets of standards, one based around HL7 / CDISC
/ FDA requirements, the other around HL7 / SNOMED CT / NHS requirements, are not
incompatible, but neither, unfortunately, are the approaches obviously similar (because in
many ways they cover different areas). A recurrent theme in developing standards for
academic trials units will therefore be trying to find ways to back both of these horses at once, and / or helping to bridge the gap between the systems.

**Change Management for Standards**

In summary, the success of implementing standards will of course depend on coherent and logical proposals for the standards themselves, but will also depend on a variety of ‘change management’ issues. In particular it will need:

a) High profile leadership and championing of the standards programme.

b) Re-iterated central support and publicity for the standards strategy.

c) Concrete backing from funding organisations, with resources for co-ordinating staff and for specific projects.

d) Tool development and dissemination, using existing expertise and developments but backing those up with central support, to reduce the costs of implementation.

e) Involvement of and consultation with key staff in the units, especially the trials unit heads and the IT leads.

f) A variety of actions to maintain a high profile for the standards programme and present it as an ongoing successful initiative

g) Demonstrable integration with other initiatives in UK academic clinical trials units.

h) Continued involvement with developing systems and standards in the NHS, such as electronic health Care Records.

i) Continued involvement with the pharmaceutical industry and the move towards standards in that sector.

j) Involvement in the wider standards debates, including membership of key Standard Development Organisations (SDOs) such as ECRIN, CDASH and BRIDG.
Recommendations 1: Clinical Databases

The data stored in trial databases could become more standardised in at least three different ways. Firstly, the same questions can be asked, i.e. the same data sought, something that is discussed below in the context of both identifiers and clinical data sets. Secondly, the questions can be standardised to have precisely the same meaning and identifiers, use the same units etc, discussed below in terms of both non-clinical and clinical data items. Thirdly the responses can be standardised and couched in terms of standard controlled vocabularies (SNOMED, MedDRA etc.).

Attempting to standardise clinical data will be an ambitious, complex project, and support from the centre will be absolutely essential if it is to succeed. Such support should take the form of ongoing management of a growing library of standardised data items, and the production of tools to help data item selection. This section provides more detail on both these requirements.

Using the Same Subject Identifiers

A fundamental aspect of standardisation is ensuring that we all identify subjects (at least those based in the UK) in the same way. In particular, the NHS Number (and its Scottish / NI equivalents) should be recognised as the central patient identifier and must be collected, for UK residents, in all trials (as it is very often now). The number is likely to become more important in the future as the main way to link clinical trial and other health data.

Ideally the NHS number would be enough to identify patients. In practice, especially for long term tracing and for matching against pathology samples, experience suggests the patient’s name is also required. Each patient’s name should be routinely collected, and stored in the format recommended by the UK Government (e-GIF standard), i.e. as title, given name, family name, initials, suffix.

Local identifiers such as Hospital Numbers have little value in linking records and their data quality is often poor. They are important, however, when communicating with centre staff using local hospital systems and so would normally be obtained.

Exact dates of birth are a useful check identifier and should be routinely collected for all subjects. Many established statistical scripts will also use them (rather than age) so that they will need to be retained in analysis data sets. The CDISC proposal to use years / months is considered unrealistic.

We should also suggest (to UKCRN) returning the NHS number along with the data in the monthly accrual returns. This would allow UKCRN to extract other demographic data about trial participants, from their own systems, without demanding too much additional work from the trials units. We do need, however, definitive clarification about the legal aspects of storing and transferring subject identifiers in this way.

If the NHS number of trial subjects was also imported back into NHS systems, the electronic health record would then include a reference to any and all trials in which an individual was participating. Over time this reference could possibly be expanded, to include (for instance) treatment group details, or a URL for unblinding, or at the least a URL / phone number for obtaining further trial details.

NHS Numbers, Hospital Numbers, Names, and Local identifiers are not required for analysis and should not appear in datasets provided for analysis. They should be stored in an approved
encrypted format. Given the limitations on many clinical DBMSs, this will mean that they will need to be stored separately from clinical data (though not necessarily on a different server). In the same way that the Joint Academic Network (JANET) provides free security certificates for web sites now, central support for obtaining and guidance for implementing database certificates should be pursued to make using encryption easier.

The recommendations relating to identifiers are summarised below:

**Rec 1.1** A core set of identifiers should be collected for all subjects in all trials: the NHS Number, their full name, a local identifier and the exact date of birth. This is to ensure correct identification and matching of subject data.

**Rec 1.2** NHS Numbers should be returned as part of the monthly accrual data.

**Rec 1.3** Subject identifiers should be in stored in an approved encrypted format.

**Common Clinical Question Sets**

Identifying core data sets for particular types of clinical data has been done in different contexts – e.g. the datasets defined in the NHS Data Dictionary\(^{10}\), and the Royal College of Pathologists’ Datasets and Tissue Pathways (for cancer)\(^{11}\), though neither of these are related specifically to clinical trials.

There is a *prima facie* case for extending the idea to trials, e.g. by disease group, or disease group / intervention type, to ensure a core of consistent data collection across a wide number of subjects. This is clearly not an IT issue *per se*: the suggestion is that the issue is raised with the UKCRN, and / or the specific Clinical Studies Groups of the topic networks, and possibly the relevant professional organisations, to test their opinion. They would need to establish the necessary item selection and management processes. The expectation would be that core datasets would start small, but gradually become more elaborate and / or more tuned to particular types of trials.

The main advantage to IT staff is that it would identify the priority areas for defining and coding questions and their responses.

**Rec 1.4** The potential for identifying minimum datasets for particular trial disease areas, or trial disease area / intervention types, should be explored with the relevant research / clinical groups.

**Common Non-Clinical Data Items**

In most trials the clinical database usually includes a set of non clinical data items, perhaps demographic data (e.g. gender, date of birth), or identifiers (e.g. NHS number, hospital number), or basic trial variables (e.g. date of registration or randomisation, treatment group, consents provided, treatment site, visit name etc.). The actual set will vary from trial to trial, but these core data items come up so often, across so many different trials, that it would seem an obvious step to define, name and code them all in the same way.

Establishing the same codes for these 12 – 20 basic data items would be relatively trivial in itself but it would help to start people thinking about standardisation, and it would be a relatively easy step to take. Making sure that the underlying definitions are consistent and that the mappings to associated systems are accurate is more important. Fortunately in this
limited area both tasks should be straightforward. Using standardised coding could only apply to new trials, but it would not be unreasonable for funders to mandate their use.

A set of relevant questions definitions already exists within the CDISC Demographic, Trial Model and Subject Characteristic domains\(^\text{12}\), and much of the same data will also be present in the NHS Data Dictionary. The default coding (e.g. NHS Data Dictionary or CDISC) is best considered on an item by item basis. Overall therefore:

**Rec 1.5** Initially, a core set of the most common non clinical items should be identified, and these should be defined consistently, and be mappable to standard codes, across all trials.

**Clinical Data – Standardising Definitions and Responses**

Several attempts have been made to construct systems for defining items and / or ‘common data elements’ or CDEs, i.e. in developing a metadata repository (MDR). The most detailed work so far has probably been that of CaBIG in the US\(^\text{13}\), though this only covers parts of the clinical domain. There are also the many thousands of entities defined within SNOMED\(^\text{14}\).

CDISC has waded in with its CDASH initiative (Clinical Data Acquisition Standards Harmonisation)\(^\text{15}\). Whilst this project is more limited in scope and focuses particularly on safety related data, it includes some features of particular relevance to clinical trials, for instance in the way it copes with (indeed encourages) repeating question groups as a pattern of data capture.

There is also the CDISC Metadata Repository (CMDR) project\(^\text{16}\), still in its relatively early stages, that takes a more rigorous approach to standardising data items, for instance by linking to CaBIG defined terminology, and which has as its scope the far more ambitious task of standardising on a set of data items that could be used, for efficacy as well as safety data, across the whole range of trial / disease areas.

The CMDR has strong backing from several large pharma companies and is likely to become a full CDISC workstream in the near future. There is general recognition amongst the supporters of the initiative that progress must be made quickly if the repository is to be useful, so some resources are likely to be made available to move this project along this year (2009).

There are other metadata repository systems being developed – e.g. for Primary Care within the ePCRN network\(^\text{17}\), within the MRC and its data sharing project\(^\text{18}\), and within the NHS itself (e.g. within NIHR).

Each of these systems could be used or contribute towards a ‘UK clinical trials MDR’. In general they are not mutually exclusive – in many cases they complement each other. For instance the question ‘adverse event experienced?’ is most easily defined and coded as a CDASH data item. The succession of responses, “Nausea”, “Alopecia”, “Thrombo-phlebitis”, etc. are best standardised using the relevant SNOMED (plus in this case MedDRA\(^\text{19}\)) codes. If they were selected from a drop down list then each available category would need to be associated with the correct code, if input as free text some form of coding process would have to occur afterwards. Similar linkages, e.g. between more specific CMDR data items and their SNOMED equivalents, are equally possible.

As illustrated in figure 2, a ‘standardised question’ can and eventually will have a variety of mappings associated with it: an alphanumeric code to use as an identifier, and ultimately
perhaps the column name in a results tabulation, probably a SNOMED digit string to link it in to that system, possibly a CMDR identifier, and potentially other relevant codes, e.g. a CaBIG CDE ID, a CDISC domain prefix, or a MedDRA, LOINC or ICD code (either directly or indirectly via pre-existing mappings), a succession of ‘key word’ tags that will allow it to be easily found, and finally one or more possible responses, each with their own code set. In some cases the response type and coding would be fixed, in others it would be user selected. Each data item would also need, in the context of the metadata repository, its own unique identifier.

In that sense ‘which is the best standard to use for data items?’ is an irrelevant question – we use all of those that are required for a particular entity. It is true, however, that different standards offer particular advantages in terms of initially identifying and defining questions and responses.

Figure 2: Possible Linkages of a Standardised Data Item

In this context the most directly relevant system appears to be the proposed CDISC MDR, because

a) it has the same purpose and scope as the MDR required for academic trials units,

b) it allows the same type of cross system linkage, e.g. to SNOMED, that we require,

c) it is – potentially at least – able to draw on a broader range of funding than many purely public sector initiatives,

d) it is in the best position to tackle integration with other CDISC systems such as CDASH, Bridg and SDTM, and

e) it is at a relatively early stage of development and therefore more open to influence and involvement than longer established schemes,

f) those involved recognise the need to rigorously define concepts and link them to controlled vocabularies, as we need, but

g) they are also aware there is a need to push on and develop a sufficiently large library of data items for the scheme to be useful in the near future.
In the circumstances the best approach for academic trials units would seem to be to join forces with the CMDR initiative – we do not want to be developing yet another metadata system ourselves if we can help it, even if we might need to manage our own instance of the repository that is generated.

Ideally we would also draw in or at least ensure inter-operability with the relevant parts of other UK MDRs – at the very least we need a mechanism to exchange information and ideas between the different groups concerned with MDRs in the UK, and eventually further afield.

The details of how collaboration / CMDR participation might be managed and funded would need to be worked out but the problems are not insoluble, indeed UKCRC already has some experience of similar situations\(^2\) – the result would be a concrete example of the UKCRC’s commitment to ‘Working in Partnership – Changing Cultures – Igniting our Potential’.

The contribution academic units can make to this originally pharma based initiative should not be under-estimated. One aspect is the work already done by different public sector groups in assembling clinical trial data items and linking them to coding systems, in particular the recent work of the ePCRN (electronic Primary Care Research Network)\(^1\). Another is the easier access we can provide to the domain experts, in different types of trials or disease areas, that will be required to ensure the validity of the data concepts being considered. Academic units, either directly or through their contacts within Trusts and Universities, are well placed to identify and organise such input.

Conversely, the academic units stand to gain a great deal from a shared initiative, not just because it will make academic and pharmaceutical trial data look more similar in the future, and collaboration therefore easier, but because it will also allow a greater range of resources, tools and expertise to be available to maintain the system.

Decisions will have to be made as to where to concentrate the initial effort in defining concepts and data items. It is suggested that the initial concentration should be on

a) Examining, comparing and where necessary converging the questions (and response sets) already worked up by the various groups involved in constructing MDRs.

b) Those questions (and response sets) which are already largely organised into standard usages, e.g. standard validated questionnaires (e.g. in quality of life, mental health), tumour staging systems, groups of blood tests.

c) Any question sets which are identified by clinical groups as being part of ‘common core’ data, i.e. that should be collected in all trials of a particular type.

It is worth noting that there is an international standard for metadata repositories (ISO 11179) that covers what metadata should be stored, how it should be organised logically, gives guidance on how terms should be defined and names derived, and stipulates how the registration process, of new metadata terms into the repository, should be managed\(^2\).

Given that a standards supporting tool should, where possible, itself be in conformance with standards, any metadata repository used for clinical trials should be ISO 11179 compliant. Note that this has most implications for those designing, documenting and managing the repository - the rather arcane details of ISO 11179 would not normally be of interest to end users. (N.B. The developing CMDR project will be ISO 11179 compliant).
If CDISC MDR is the best, though certainly not the only, starting point in terms of defining questions, then as far as defining and coding possible responses are concerned the prime source for standardisation and coding is already obvious - it has to be SNOMED CT, because of its planned use throughout the NHS,14.

The SNOMED link is vital if we are to realise the potential benefits of the Research Capability Programme – the more clinical trial data is couched, stored and analysed in the same terms as routine clinical data, the more we can use and combine data from a variety of sources. Trials IT staff will, however, need to understand how SNOMED is structured and become familiar with the main elements of the system.

Other codes should be added where relevant – e.g. MedDRA19 will also be needed for coding AEs / SAEs, ICD1022 would be useful for causes of death. However, rather than add these additional codes to the responses directly, it will often be more sensible to use external cross-system mappings, to minimise the impact of different versions of the standards being introduced at various time points.

This approach to standardising clinical data is unashamedly pragmatic, and tries to make use of the best systems and approaches that are already in existence. In essence it tries to bolt together the pharmaceutical’s industry’s developing CDISC based systems for defining and storing clinical trial questions, with the NHS’s (and others’) developing SNOMED based systems for storing clinical data responses and values. Because of that it may appear rather messy. It is preferable, however, to trying to construct an elaborate, all encompassing, and expensive single scheme from scratch, and in any case, the CDISC MDR is already seeking a similar linkage. (One of the objectives of the CMDR pilot is to show “how terminologies such as MEDDRA, SNOMED, LOINC can be linked together” ... with the CMDR system)23. In summary, the recommendations relating to standardising clinical data are:

**Rec 1.6** We need to identify, share and use standardised data items, each of which is clearly defined, has links to one or more standardised (coded) responses, and links as required to other standard systems.

**Rec 1.7** The most effective and cheapest way of developing and resourcing the necessary collection of data items – a metadata repository – would be to collaborate within the CDISC MDR project currently being developed for the pharmaceutical industry.

**Rec 1.8** The initial sources for the data items should be (i) the questions already developed under the CMDR initiative, (ii) work already done by trialists and other public sector groups in developing standard question sets, (iii) pre-existing standard instruments and questionnaires (iv) data items identified as ‘core’ for particular trial types.

**Rec 1.9** The standard for coding data item responses must be SNOMED CT. Other coding schemes should be incorporated if relevant, either directly or via external mapping schemes.

**Rec 1.10** In any case the metadata repository should be ISO 11179 compliant.

**Rec 1.11** Trials IT (and other) staff need to become familiar with the SNOMED CT system, and assess and explore how it can best be used in the future.
The Need for a Managed Data Item Library and a Selection Tool

The difficulty is, of course, that no-one but the keenest evangelist has the time or inclination to identify, define and link sets of standard data items or, once those items are established, to search for the correct data item to use in any particular context, and then manually check and transfer the details to a list of data items for a trial. A trip to the CaBIG site to browse common data elements (very, very slowly) only underlines the total impracticality of such an approach.

It is therefore absolutely essential, if standards are to be successfully introduced into clinical data, that

a) Trials units have access to a library of standardised data items, maintained centrally, containing within it the pre-defined, pre-linked, and pre-coded items of the metadata repository. Such a library may be ‘the’ repository, or it may be a UK only, or UK academic trials unit ‘mirror’ of the main system (the latter would probably be easier to manage and allow candidates for inclusion to be considered more quickly).

b) An easy to use ‘data item selection’ tool is developed, so that those designating data items for a trial can easily find the relevant items and automatically drag them to their correct place in the specification’s list.

These proposals are quite ambitious, and they will involve an ongoing cost to the centre, but they are at the heart of standardising the clinical data structures in clinical trials. The greater the utility of the library and related tools, the greater the number of potential users, and thus the wider the potential source of funds. In particular if the library and client tools, like the development of the metadata repository itself, can be shared with UK based pharma, the opportunity arises not just to share costs but also to develop an innovative facility that would be of lasting value and of interest internationally.

*Figure 3: Using a common library of data items*

As figure 3 illustrates, the idea is that in the trials units a local library of data items would be used, for maximum flexibility and innovation. This would be initially populated and updated by the central managed store, but allow additional items to be defined locally. These new items could then be uploaded for consideration of inclusion centrally.
The input to the tool would be the trial protocol, the trial schedule (i.e. visits and forms structure), preferably as a standardised XML file, the local repository of data items and of course the decisions of the relevant staff.

The output would not be a database and/or a set of CRFs, whether paper or electronic. It would instead be a detailed XML metadata file (based on the CDISC ODM schema), that can then be fed into automatic or manual database and CRF generation tools, and which also acts as a repository of the trial’s metadata for other purposes. In that way the trial design process is removed from any consideration of the particular database and/or data collection system to be used and the tool becomes generic.

The users in the trials unit therefore need access to a module that can:

a) make reviewing, selecting and updating data items easy for users,

b) manage the updating of the local repository,

c) return requests for inclusion of new data items

d) easily incorporate range and logic checks, skips and other extra features

e) incorporate multi-lingual and multiple character set questions

f) document the trial design process,

g) act as a vehicle for change management, and

h) generate the XML outputs necessary for the next stages of database/CRF construction.

The aim would be to develop (or buy) a tool that met this functional specification, though not necessarily all components of that specification in the first instance.

In fact tools are already being developed that fit many of the criteria listed. The ePCRN group, for instance, based in the Birmingham Primary Care Trials Unit, has created a system that allows users to browse and/or search for a store of (pre-defined, pre-linked) data items, obtain a clear definition of each term, and then transfer a selected item to a CRF under construction. A few miles away, the University of Warwick’s Trials Unit already has a suite of XML tools that can take an XML based metadata description, and use it to create both the database, the web pages for the eCRFs, and the intervening programme code, all in the space of a few seconds. Several other units have developed or are developing their own local repositories of data items.

One of the most advanced of the current projects is the metadata aware trial design system constructed as part of the CancerGrid project by Oxford University’s Computing Labs. This makes use of an ISO11179 compliant metadata registry to allow users to select and review data items, and also—already—has the capacity to link through to a central data store (in this case at CaBIG) to view existing terms and propose new, locally defined terms. Though it has been developed in the UK, the metadata repository function is currently being evaluated by the Mayo Clinic in the US. Note that the metadata registry is available free to download.
At the centre, any metadata library would probably require only a few core staff, but would need to be able to call on external expertise as required. It is expected that trials units would want to put forward their own candidate data items for inclusion in the standardised set. Reviewing such items and their associated responses, resolving ambiguities, avoiding duplications, ‘laying off’ difficult questions for scientific / medical advice would all be amongst the functions carried out by the central facility. The centre could also co-ordinate the production of relevant support materials and training.

To facilitate maintenance of the central library, and also to help support collaborative development, there might be advantages in making this a web based system, with trials units having controlled access to only their own trial definitions (though this would need users to be completely satisfied that security measures were effective).

The proposals are neither technically outlandish nor out of step with developments already taking place. More than one project is ‘almost there’. The difference is in the scale and ambition of the proposal, i.e. the intention to try and join together the various groups and systems and use the best features of each. By its very nature, however, any shared repository becomes more useful the more inclusive it becomes. In summary therefore:

**Rec 1.12** A central library of standardised data items should be established, run and funded centrally, linked to local item databases and acting as the master repository for them. Such a library might be ‘the’ repository, or it might be a UK academic trials unit ‘mirror’ of the main system.

**Rec 1.13** A Data item specification tool should be developed to support local access to a data item library, updating that library from a central store, and to provide documentary support for trial design and amendment

**Rec 1.14** Appropriate Materials and Training should be developed and provided to ensure all relevant staff understand the aims and details of the data item library project, and know of the tools available to support it.

**Recommendations 2: Trial Description and Metadata**

**Trial Descriptions**

There are a variety of standard datasets focused on the trial as a whole. Some are relatively simple, like registration data requirements (now in the process of being included in the BRIDG model) or the Consort dataset established by medical journal editors.

The BRIDG model of ‘protocol driven activity’\(^{26}\), developed by a consortium of HL7, CDISC, and CaBiG, is the emerging general model for a trial and any standards in this area will need, ultimately, to seek inter-operability with it. It is only partially developed however and, being based on the HL7 RIM (Reference Information Model)\(^ {27}\), is relatively complex.

Of most immediate relevance to academic trials units is probably the data used within the NIHR’s longitudinal research record (LRR)\(^ {28}\), not least because this data will be bound up with the IRAS\(^{29}\) and CSP\(^{30}\) processes, so most units will need to have and deal with this data set in any case.
The LRR dataset and model should therefore be at the core of ‘whole trial description’ data. Some of the major fields involved are listed in the details given for the Research Project Class in the NIHR’s Information Systems Enterprise Architecture paper [ref].

There will need to be a checking, mapping and extension process applied to this core, probably best done under the aegis of the NIHR, to ensure that all relevant data points from related data sets are covered in the final model. The current largely static model will also have to be extended to include time related data, recording important trial milestones.

Like the clinical data items discussed above, it makes sense for the trial model to have its metadata stored in a metadata registry, so that terms and the relationships between them are defined without ambiguity and can be used consistently and with confidence. There is no reason why the same metadata repository that needs developing for clinical data items can not also hold this trial related data.

The goal is a comprehensive, shared characterisation of a trial, for use within the trials unit itself, for internal use within NIHR, as a source of data displayed on relevant web pages, as a target for resource discovery systems (e.g. NCRI’s ONIX system\(^{31}\)), and as a single source for reports and data extraction — e.g. for uploading to registration systems, or providing feedback to funders. Ideally, if the appropriate links, schemas and file sharing systems are in place, this data should only need entering once, however many different organisations need to access it. Such a pool of shared data could also further streamline the IRAS process.

The final trial description model will need to be integrated into a wider model of general trial administration (see recommendations 4.12 – 4.14) and indeed could be seen as the first major component of that model.

An over arching XML schema will need to be established so that trial details can be stored, discovered and distributed in a systematic way, whatever the nature of the source systems. Related tools, e.g. XSL transforms, will be required to support the model’s dissemination and use. Descriptions, explanations and training materials will also be useful in making the trial description standard better understood and adopted.

Convergence with the relevant parts of the BRIDG system should be sought, both during initial model development and afterwards. It will also be important to try to influence the development of BRIDG to reflect the needs of the UK academic trials community, in this context and others, so we should seek long term involvement in this international project.

BRIDG has put much effort into developing structured versions of protocols\(^{32}\), which potentially allows the scientific content of trials (rationale, inclusion and exclusion criteria, analysis plans and powering, etc.) to be included in databases, over and above simple administrative fields. We need to assess the applicability and practicality of this work for inclusion in the trial description model.

The recommendations for a trial description standard are therefore

Rec 2.1 NIHR needs to co-ordinate the development of a shared data model for characterising trials in a standard way, based around but extending the NIHR’s Longitudinal Research Record with data points from other systems.
**Rec 2.2** XML schemas, tools and explanatory material, etc. need to be developed to support the dissemination and uptake of the resultant model.

**Rec 2.3** In the longer term we should seek convergence with and / or influence the relevant parts of the BRIDG model. This may allow the base model to be extended to include more details on the scientific content of trials.

**Schedule Descriptions**

For simple trials the ‘schedule’ – the pattern of data collection time points (‘visits’) and data collection instruments (‘forms’) – may be a trivial exercise: all subjects may experience the same data capture routine, as a series of visits dated from randomisation.

For many trials, however, this process can be complex, for instance for cross over designs, or when different groups of visits are dated from different subject milestones (date of surgery, end of radiotherapy, key results obtained etc). In addition different visit schedules and forms may apply to for different subjects, e.g. depending on their treatment group, or, if the schedule is revised mid-way, when they entered the trial.

Carefully mapping out the trial schedule(s) is an important component of the trial’s overall description, but it is also vital as an initial step in designing the data capture instruments and database, and will also be used to structure future workflow and identify which data is late arriving. Having a standard way of doing this is therefore important if we are ever to develop flexible, modular, standards driven processes for trial database construction.

The best known model or standard that deals extensively with this area is the BRIDG model, though the part dealing with trial schedules is based on earlier CDISC components. The BRIDG model is a UML based collection of classes rather than a schema, so it will be necessary to develop, or preferably borrow / adapt, a suitable schema to describe trial structures in a computable form.

There is also the PCROM (Primary Care Research Object Model\(^{33}\)), developed by the ePCRN. This model was developed to meet the specific needs of research in primary care (and like BRIDG covers much more than just the trial’s schedule) but it is cognizant of and comparable with the general BRIDG scheme, and is now seeking harmonisation with it. We should be able to benefit from the experience and insights gained in developing the PCROM in seeking to develop a schedule schema from the BRIDG model.

Exactly who should co-ordinate the development of such a schema will depend partly on what expertise and related tools already exists, and where, but as stated previously ISWG involvement in the way the BRIDG model develops will be important to maximise its usefulness to UK academic units.

As with most standards, tools and support materials will also need to be developed, to help clarify the nature of the standard, to help generate / describe schedules in the appropriate formats, and to support the spread and uptake of the schema. The relevant recommendations are therefore

**Rec 2.4** An XML schema needs to be developed to capture trial Schedule details – visits, forms, triggers, segments etc – based on the BRIDG model.
Rec 2.5 Tools need to be developed that allows users to easily define trial structural/schedule components, generating the output in a suitable XML file.

Question Descriptions

A standard for describing how items are assembled, showing their placement, order and the relationships between them, i.e. the detailed metadata of trial data stores, is different from standards for the data item themselves (as discussed in the context of standards for clinical databases) though it is clearly related and can often be generated from the same tools. In fact this type of data exists at three steadily more comprehensive levels:

a) database metadata, which includes only the data items, response categories, and their ordering and placement.

b) question metadata, which includes the detailed caption, validation logic, skipping and messaging to be found in the system, as well as the database metadata, and

c) presentation metadata, which includes the design details, placement on screen/page of the various elements, including headings, lines and graphics, as well as all the question metadata.

We need to develop a standard ways of describing each of these datasets. While the BRIDG (and PCROM) classes model the trial itself – e.g. interventions and assessments – they do not span the structure of the supporting database, let alone the way questions are presented to the user. Instead it is the CDISC operational data model (ODM) that, so far, has provided the most complete model of trial databases.

The ODM model covers most, but not all, aspects of a database’s structure (e.g. full validation and skipping logic are missing). It would therefore need extending to fully represent question and presentation metadata, but that would be a relatively straightforward exercise. For the foreseeable future therefore, despite some questions over its long term future (compared to HL7 based structures) it appears to be the most appropriate standard to use in this context.

Each of the metadata types listed above are a by product of the relevant design and later modification processes. Their generation, using the appropriate XML schema, should therefore be included as part of the requirement specification of the relevant tools.

Establishing the schemas and tools would allow metadata to be generated for new trials in a relatively straightforward fashion, and thence exported, e.g. as part of a collaboration, or used to support other activities, for instance to provide a list of validation checks to support database validation.

The real problem would be all the data stored in ‘legacy’ systems, i.e. those for which metadata does not currently exist in the recommended XML schema, which of course includes almost all current trials! To be useful within a reasonable time frame tools will need to be developed to allow existing systems, at least those in the major formats, to be mapped back to these metadata standards (and those for the trial schedule as well).

The time and costs involved in cataloguing very old trial metadata rigorously should not be underestimated, given the non standardised nature of the data dictionaries available for some older trials; it may be that a ‘best in the circumstances’ health warning will need to be applied to the results (which itself only underlines the importance of standardising on data items as...
soon as possible). In time, however, the process would allow the metadata of most existing and new trials to be brought together (e.g. linked to the NIHR portfolio database) thereby providing a detailed record of trial activity and available data items.

The recommendations relating to metadata standards are therefore:

Rec 2.6 XML schemas needs to be developed to capture trial metadata details – questions, formats, associated logic, captions etc. – based on but extending where necessary the CDISC ODM model.

Rec 2.7 Where the ability to extract metadata in these standard formats does not exist, which is the case for almost all existing systems, suitable tools will need to be developed, at least for the major systems currently in use.

Recommendations 3: File Management

This section considers the potential standards in the way clinical data files are structured, and thereby transferred from one context to the other. There are three principal opportunities for standardisation:

a) the way data is structured when it is initially captured and transferred to the primary database,

b) the ways in which data is structured when it is transferred from one organisation to another, and

c) the way in which data (and metadata) is structured when it is archived for long term curation.

Data Collection Formats

Within clinical trial systems there are no general standards for structuring the data received from eCRFs – they tend to be vendor dependent and tightly tied to the architecture of the database and / or collection instrument, often with a simple correspondence between on screen text box and database field.

That is not normally a problem, though it does bind the whole process together into one indivisible lump – tying together the construction and distribution of the data collection instrument with the capture process and subsequent transfer to the database. If the data capture process can instead be dislocated, by the insertion of an intermediate XML based data format for the returned data, there is the opportunity to modularise the process, with consequent gains in flexibility, responsiveness and competition.

Such a process may become more critical in the future, if – as seems very likely – other data sources are used for trial data: spreadsheets or machine data from laboratories, downloads from national disease centres (e.g. the NCIN) or the Secondary User Service, even electronic health care records (eHCRs) from NHS operational records. Figure 4 illustrates some of the variety that might face trials units.

We will need, in a few years time if not immediately, to develop a standard to act as the ‘final common pathway’ for data capture, an XML format that all incoming data can be transformed
into, leaving only a single step to squirrel it away in the database. Once established the systems and organisations that generate data then have a common target, whether they manufacture radiotherapy machines, publish laboratory information systems, collect hospital episode statistics, or design eCRF tools.

![Diagram of data entry into the database]

**Figure 4**: Potential future routes for data collection

The most obvious basis for such a standard is HL7’s Clinical Document Architecture\(^{36}\) (CDA), partly because this is the HL7 instrument designed to capture any permanent health related record, and a CRF is certainly one of those, but mostly because HL7 CDA is the schema used by the systems currently being introduced to implement eHCRs in the NHS.

CDA is complex, and there appears to be some debate about the best way to implement it within NHS systems. Nevertheless it is a comprehensive and very flexible method of capturing data that can handle code mapping if necessary (as well as free text and images).

The best approach would probably be to develop a common standard for clinical trial data that was a simplified version of CDA, i.e. something that operational records could be converted into using XSL transforms. It will first be necessary, however, to develop that standard, working in conjunction with those involved with CDA in the NHS. The recommendation are therefore:

**Rec 3.1** A group should be established, involving amongst others, ISWG members, CfH staff, and staff from pharma, to investigate and try to develop a common XML schema for data destined for clinical trial systems, based on HL7 CDA.
Rec 3.2 In the longer term, data should be captured and returned to trials units using XML and the CDA based schema derived by the group above, rather than in a proprietary database format.

Collaboration Formats

Traditionally data has been transferred between organisations using plain vanilla formats like CSV, Excel sheets or SAS transport files. This is adequate in many situations, particularly where the structure of the data is well understood by the recipient, but it normally removes audit and other supplementary data, and lacks the flexibility of an XML based approach.

There are two main modern standards designed for trial data transfer, both developed by CDISC.

a) The operational data model (ODM) is the most flexible, is relatively easy to understand and use, can mirror exactly the structure of the original data, and can include audit trial and electronic signature data. The problem is that it can generate enormous, unwieldy, XML files which can be difficult to manage.

There are also some question marks over the long term future of the ODM35, though for the moment there is no viable alternative for transforming the full set of original clinical trial data into XML.

b) The CDISC Standard Data Tabular Model, or SDTM37, involves a restructuring of data according to a predefined set of rules and conventions (though the details will vary from trial to trial, the overall structure remains the same).

It generates database tables rather than XML files, and is also a useful ‘half-way house’ for further data processing, e.g. into analysis data sets or for reporting.

In effect a conversion to SDTM ‘chunks’ the primary data into tables for each major data area (‘domains’ in CDISC terms), with a record for each instance of each ‘base entity’. Examples of domains are adverse events, concomitant medication, vital signs, laboratory results and exposure (i.e. study treatment).

As illustrated by figure 5, a domain record has initial columns that identify the study, subject, and visit, followed by domain specific fields, each with the same two letter prefix, that carry the record’s data (an actual CM table would have many more fields). SDTM includes a useful coding convention for column names, and provides a powerful way of re-organising data into a very similar format, whatever the source trial.

The difficulty with SDTM is the cost of mapping the usually monolithic structure of the primary clinical database to the SDTM model and then accurately transforming the data (and documenting the final structure in so called “define.xml” files). The pharmaceutical industry, faced with requests from the FDA to use CDISC for submission data, is finding this expensive enough – it will probably be outside the budgets of academic units unless much of the mapping can be done prospectively, at the trial construction stage.
<table>
<thead>
<tr>
<th>STUDYID</th>
<th>DOMAIN</th>
<th>USUBJID</th>
<th>CMSEQ</th>
<th>CMTRT</th>
<th>CMDOSFRQ</th>
<th>CMSTDT</th>
<th>CMENDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>1</td>
<td>1</td>
<td>Aspirin PRN</td>
<td>2003-10-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>1</td>
<td>2</td>
<td>Propanol BID</td>
<td>2004-01-01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>1</td>
<td>3</td>
<td>Chlorpromazine BID</td>
<td>2004-01-01</td>
<td>2004-03-23</td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>1</td>
<td>4</td>
<td>Spiranalactone BID</td>
<td>2004-02-08</td>
<td>2004-02-14</td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>1</td>
<td>5</td>
<td>Feroniside BID</td>
<td>2004-02-08</td>
<td>2004-01-25</td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>1</td>
<td>6</td>
<td>St John's Wort PRN</td>
<td>2004-02-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>2</td>
<td>1</td>
<td>Diazepam BID</td>
<td>2004-01-01</td>
<td>2004-02-14</td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>2</td>
<td>2</td>
<td>Aspirin BID</td>
<td>2004-02-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>2</td>
<td>3</td>
<td>Tri-Glycerol Nitate ONCE</td>
<td>2004-02-08</td>
<td>2004-02-08</td>
<td></td>
</tr>
<tr>
<td>ABC123</td>
<td>CM</td>
<td>3</td>
<td>1</td>
<td>Warfarin Daily</td>
<td>2001-04-01</td>
<td>2001-11-01</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: A small portion of a CDISC SDTM file (Concomitant Medication domain)

Despite these problems, SDTM is not only useful if sharing data with a pharmaceutical collaborator, it is also intrinsically valuable as a way of processing data, especially in large trials, and comparing and combining data from different sources (e.g. in meta-analyses). For these reasons the standard should become better known within the academic trials community and trial design processes should support it as much as possible (N.B. CDASH is clearly designed to support later transformation to SDTM).

For instance the standardised data items discussed in the section on clinical databases could usefully link to a domain reference and even the name of the CDISC field that they would normally be transformed into. For most academic units, especially the smaller ones, prospectively linking data items to SDTM in this way would be the only practical way to approach the transform process – trying to do it retrospectively would be far too expensive. It may be, however, that we can beg, borrow or steal expertise gained in this area by the pharmaceutical industry. The recommendations relating to standards for data transfer are therefore:

**Rec 3.3** Trials units should become more familiar with the CDISC SDTM model as the preferred vehicle, in the longer term, for file transfer as well as submission to regulatory authorities.

**Rec 3.4** We should acquire and/or develop mechanisms for turning clinical data into SDTM Datasets, possibly with the help of partners in the pharmaceutical industry.

Another CDISC group is working up the Analysis Data Model (ADAM), which is an attempt to standardise how derived and statistical data is presented (chiefly to FDA reviewers though it could have potentially wider benefits if it was used more widely, e.g. for peer review of results).

The model is at the moment relatively undeveloped, and implementing standards in this area appears to have a relatively low priority. It might be useful in the future, however, to at least investigate the need for standardised approaches to producing Analysis Datasets.]
Long Term Curation of Data

Over and above the demands on the physical systems and substrates used for long term storage, the data itself has to be structured in a way which is (ideally) simple, non-proprietary and self describing. Only that way can it can be guaranteed that after twenty or thirty years (say) data files will still be readable by whatever systems are around at the time.

This clearly means XML, and for trials the only schema that can currently capture all of a study’s data, including audit information, is the CDISC operational data model (ODM). The ODM also has the advantage that the metadata can be stored in the same format.

The problems of large files can be partially solved by compression (as long as the longevity of the algorithm can be guaranteed) though the continuously dropping costs of storage capacity mean that this issue becomes much less important over time. The problem of deprecation of the ODM in the future, and its possible replacement by an HL7 based system, can probably (if need be) be solved by application of the appropriate XSL transform.

If we are to begin systematic curation of data now, as opposed to 5 or 10 years time, then mechanisms will also need to be developed that allow the extraction of data from at least the major existing systems (newer systems ought to have a ‘archive module’ built in). The issue is an extension of the problem of capturing metadata from existing systems – here both data and metadata are required, though in both cases the ODM can be used.

Developing the necessary tools will require central co-ordination and support. It would also seem sensible to centralise the curation process itself, because of the specialist kit, security and expertise required. The result would be a national repository of trial data – a valuable resource in its own right as well as being an example of innovative research support.

The recommendations are therefore:

**Rec 3.5** The CDISC Operational Data Model (ODM), or where necessary a simple extension of it, should be used as the data format for trial data going into long term curation.

**Rec 3.6** Mechanisms for extracting data and metadata for curation purposes from existing systems should be developed, with central support.

**Rec 3.7** A national repository of clinical trial data be established to manage the long term curation of trial data.
Recommendations 4: Trial Management and Administration

Trial management covers a variety of activities, three of which are obvious candidates for the greater use of data standards.

a) Pharmacovigilance and safety reporting, not least because a de facto data standard already exists in MedDRA.

b) Contact data management, because standardising on people and organisation data would allow that data to be shared, managed and updated much more efficiently.

c) General trial administration activities, such as managing and monitoring sites, tracking data receipt, query management, though a generally accepted model of trial administration would be required beforehand.

Each of these are considered below.

Pharmacovigilance

The use of MedDRA for SAE reporting is mandated\(^ {39}\), so this is one of the few areas where we already have a clear data standard. There are also relevant process standards (for evaluation time frames, line listings, SUSAR reporting etc.). We should therefore concentrate on developing systems to better support the deployment and use of MedDRA and these processes.

Some units already use MedDRA but many do not, and there has often been confusion about the terms and conditions under which a subscription can be purchased. It ought to be possible to cut through this confusion, ensure future coverage for all trials units, and seek some economies of scale, by negotiating a central MedDRA licence for academic trials units.

MedDRA is relatively simple to code against but because the system is ‘multi-axial’ there is often a choice as to exactly how symptoms are grouped and sub-totalled – the trick is therefore in maintaining consistency in its use. It also helps if there are consistent approaches to versioning (MedDRA is revised every six months), and to working with Systematised Medical Queries\(^ {40}\), or SMQs, which are precompiled signal detectors within the system.

Given that all units dealing with IMP trials have essentially the same requirements in terms of expedited SAE and SUSAR reporting, it would be absurd for every unit to develop essentially the same system in parallel (though that is exactly what is beginning to happen). There is clear potential for a generic system, possibly based on the best of the existing systems, which (for instance) could make use of the NIHR portal to manage investigators’ evaluations, rather than the cumbersome fax based methods most units use now.

Such a system would also need a local component, partly because in many trials SAEs also contain study specific information, partly because SAEs need reconciliation with non expedited adverse event reporting. Once the information is stored and processed locally, however, one can envisage a system that uploads the core data to a central repository, where it can be updated remotely as necessary. The central system can then be interrogated for reports, manage electronic submission to MHRA and EUDRAVigilance and possibly also orchestrate the coding.
MedDRA can be used to code all adverse events, not just SAEs. This raises the possibility of developing a generic MedDRA coding system for batches of AEs, especially those reported as free text, as well as pre-linking categorised data items to MedDRA codes, and / or of mapping SNOMED codes to MedDRA. (Future NHS reporting of AEs may well be expressed natively in SNOMED, so some form of mapping is going to be almost inevitable).

Developing or buying general AE coding systems would depend on demand, but would seem a useful and natural extension to the SAE based system. Such a system could be mostly local, with periodic download, or possibly use web services and a central repository. Again existing expertise and initiatives should be used where possible.

Current training in MedDRA is usually very expensive. Based on common policies and systems, it should be possible for the centre to produce appropriate support and training materials as well as providing, or negotiating, general MedDRA training at a better rate.

The pharmacovigilance related recommendations are therefore:

**Rec 4.1** Negotiate and fund centrally a single MedDRA subscription arrangement for all academic trials units.

**Rec 4.2** For ISWG to liaise with the other UKCRN working groups, and any other relevant organisations, to develop a consistent MedDRA usage and coding strategy.

**Rec 4.3** For existing coding systems and tools to be examined, and decisions taken on a) a requirements specification and b) a development and / or procurement strategy for a national pharmacovigilance system (for SAEs at least)

**Rec 4.4** If there is a demand, to also specify and plan the development of MedDRA coding systems for all AEs, especially those reported as free text.

**Rec 4.5** For the development of the proposed pharmacovigilance system(s) to be costed and planned as a centrally supported project, with appropriate long term support and maintenance.

**Rec 4.6** The development or purchase of appropriate support materials and training packages in the use of the pharmacovigilance systems and tools described above, in MedDRA coding and in the use of SMQs.

**Contact Data**

At the moment we have trials units, trust research departments and the centre (UKCRN, NIHR etc.) all trying to organise their own pools of contact data – lists of people and organisations and the relationships between them. In many cases this data is not even integrated at the level of the trials unit – individual managers have their own spreadsheets and databases. The data in most of these systems is inevitably out of data and inconsistent, is usually incomplete, and very often includes spurious entries and duplications. The total time spent trying to manage this mess is absurd.

A coherent managed pool of contact data and identifiers is possible, particularly if we can piggy back on the work being done within NIHR on an identity management scheme for (potentially) all NHS staff and health researchers, and the location identifiers managed by the NHS’s Organisation Data Service (ODS). Such a scheme would see local caches of contact data
derived from central datasets, those caches returning updated data and new candidate entries after they had been entered locally.

The idea would be to couple flexibility for users at the satellite sites, and efficient update by those most familiar with the data, with control, validation and management centrally. For it to work, however, we will first have to standardise on how contact data is structured. Though there are no universally accepted schemas for people and organisation data, in this context it would make little sense to use anything other than the relevant parts of the NIHR and ODS data models (and, in terms of the detailed data structure, the UK government e-GIF standards).

It may be important in the longer term to try and help ensure interoperability between these standards and those that emerge from the BRIDG trial model. Given both the NHS and BRIDG claim adherence to HL7 this ought to be possible.

As an initial step, trials units and other consumers of contact data need to become familiar with these standards and be encouraged to transform their own data into this format, or at least include it as a subset of their own data models.

Clearly not all people / places of concern to clinical trials units are UK based, or would otherwise be part of the NIHR scheme, so simply using the NHS datasets will be insufficient. We need to find ways of safely extending them.

It will also be necessary to clarify the legal implications of sharing contact data in this way and where necessary put in place appropriate security measures, permission management systems etc.

The work involved in the proposed scheme is not underestimated, but the potential savings in accuracy and efficiency are enormous, and much of what is proposed is an extension of what is happening in any case. The recommendations relating to contact data are therefore:

**Rec 4.7** Underlying data models and structures for person and location records should conform to NIHR, ODS and UK e-GIF standards, and trials units supported in making any necessary changes to their data.

**Rec 4.8** The data in the identity management scheme proposed by NIHR and the location codes in ODS should be made available to trials units, giving them access to both accurate data and unique, managed identifiers.

**Rec 4.9** Local repositories of organisation and people data should be established and maintained by each trials unit, linked to the central ‘master’ data store. Various applications could be developed to work with such a data store.

**Rec 4.10** A client tool needs to be developed and made freely available to allow the managed download / upload process to be made as transparent as possible to end users. One or more existing contact systems may be able to provide the basis for such a tool.

**Rec 4.11** The legal boundaries of information sharing of this type need investigation and the proper safeguards and systems put in place.
General Trial Administration

Trial administration systems tend to be piecemeal, often organised on a trial by trial basis, and bolted on as an afterthought to the central task of capturing and storing the clinical data. Developing comprehensive, integrated trial administration systems (linking to, for instance, stores of people and organisation data) can bring substantial benefits to trials units, e.g. in terms of tracking data receipt, managing site agreements, and providing management reports.

If the data is structured in a standardised way, however, the benefits can go much wider. It can be more easily shared with the centre (especially if some of the coding was originally derived from there), e.g. when setting up trials, updating trial status and accrual figures, providing details of new centres etc. If the structure is also shared with trust based R&D systems then much of it can potentially be updated ‘at source’, e.g. delegation log data and training records, and then electronically transferred to unit systems.

The problem is that there is as yet no widely accepted or comprehensive model for trial and site administration (The BRIDG trial model currently concentrates on the trial itself and its protocol and interventions rather than its administration). Instead various organisations are currently producing their own models of trial administration. These include NIHR and those developing Trust / Network systems like EDGE, as well as those units who are building integrated administration systems.

It is important that these efforts are co-ordinated, so that the models start to converge. The wider trials IT community should also be involved, with the model shared and discussed as it evolves. In addition, we need to ensure that the model being developed by the UK academic clinical trials community is fed back into groups like BRIDG, so that it remains aligned to the more general schemes being developed.

Once a common data model has been developed it becomes possible to consider how it can best be used, both within trials units and between those units and other organisations such as NIHR. This could mean not just clarifying requirements, but also – for instance – reviewing current systems and best practice, developing linking tools, user interfaces and standard reports. Prioritisation of tasks can then occur, leading to adoption, development or procurement of the most appropriate systems.

Rec 4.12 A common data model should be developed for UK trial administration, the development process involving all relevant parties and sharing thoughts and interim results with the wider trials IT community.

Rec 4.13 The UK trials IT community should be represented on BRIDG (directly or indirectly) and other groups involved in developing these models.

Rec 4.14 Once a common data model has been established, a requirements gathering exercise should be followed by an assessment of the feasibility of developing or purchasing trial administration systems, building on the work already done in some units.
Time Frames

The time that will be needed to implement the various recommendations in this report, or at least those that are accepted as reasonable, will obviously depend upon the resources made available and the priorities of the stakeholder group (e.g. a reconstituted DIMS board) that retains oversight of the process.

Nevertheless this section attempts – in very broad terms – to map out what could be expected when, assuming some central staff to drive the process forward, co-ordinate the various team efforts as well as provide direct input, and document and disseminate the standards and systems that result.

Figure 6: Time frames for developing standards in question sets

In general, the expectation is that specifying or extending most standards, e.g. the definitions of XML schemas, can be accomplished within three years, and much less in many cases. What may take longer in some cases is the full development of systems around those standards.

For instance, figure 6 indicates a possible time frame for moving towards using common question sets in trials. The idea of identifying common non clinical data items in the same way is neither technically nor organisationally difficult and ought to be accomplished within a few months. The trick will be to push the change through, which is why the support of funders becomes important.

The idea of using a consistent set of subject identifiers centred around the NHS number should also be straightforward, but there are legal issues that must be properly clarified first. Trials units need to have their say about how this proposal might affect them, but with luck the proposed subject identifier regime could be introduced within 12 months.

By that time too central accrual monitoring should have been amended to include the NHS number, and consideration can turn to the feasibility and costs of ‘backfilling’ missing identifiers for older trials. What happens next (say in year 3) will depend upon those costs. Also in this year negotiations would start with eHCR providers to investigate how trial
membership can best be fed back into care records. Further development and integration of identifiers may depend on the evolution of randomisation systems and trust based research administration systems, as well as the care record systems themselves.

Figure 7 illustrates a possible time frame for developing standard data items. It will be necessary to start the development work on these three inter-related and vital projects in the first year: examining existing work in this area, establishing a consensus on the overall strategy, developing a clear specification for local and central systems, costing the whole exercise etc.

![Diagram of time frames for developing standardised data items and related systems]

**Figure 7**: Time frames for developing standardised data items and related systems

After that a period of development is envisaged for at least a further year, with initial dissemination of systems not occurring until year 3. After that the growth and maintenance of the library, and further review and development of related systems, is seen as ongoing.

Figure 8 shows anticipated time scales for developing the XML schemas involved with trial, schedule and data item description. The time consuming elements here are less the development of the standards, which should take less than 2 years, as converging them with other related standards – in particular the BRIDG model – and developing related systems to output data in the specified format.
Figure 8: Time frames for developing trial description standards

Figure 9: Time frames for developing file management standards

Figure 9 shows the proposed time scale for developing file management standards. That for long term curation is linked to those for metadata production, whilst the collaboration standard (SDTM) is already with us. The need to develop a data collection format is probably not urgent at this time, but should begin in earnest in year 2, with a view to publishing it at the end of year 3.
Figure 10 illustrates a proposed time scale for trial administration standards. Using MedDRA for pharmacovigilance seems inevitable, so the initial effort is in supporting that with a better licence deal and consistent policies. After that attention can turn to developing a national system for pharmacovigilance reporting.

Progress on rationalising contact data depends partly on progress elsewhere (e.g. in the NIHR’s identity management programme) but within the three year time period one would expect initial deployment and testing of a linked contacts management system.

![Diagram of time frames for developing trial administration standards]

**Figure 10:** Time frames for developing trial administration standards

Developing a general trial administration model (linked to the work on trial description and contact management) may be quite time consuming because of the wide scope of the model and the numbers of people and organisations already involved in this activity. It should be possible, however, to establish and start testing a schema by the end of year 3.

**System Vendors and Time Frames**

One difficulty with simply stipulating standards is that system vendors need to support them if they are to appear in commercial systems – and some trials units are going to want to continue to purchase systems (or have that decision taken for them by a parent organisation) whatever the merits others may see in developing systems in house.

For that reason it will be important to try and involve system vendors in the standards development process, something they have shown themselves willing and able to do in the past (e.g. CDISC includes many software vendors amongst its sponsors). If we are serious about developing and using standards, however, we – UK the academic trials community, with as many pharmaceutical partners as we can muster in this context – have to make it clear that at some point (5 years from the inception of the process does not seem unreasonable) we will be looking for support for these standards in the products we purchase.
Figure 11: Time frames for developing standards and systems vendors / developers

For that reason it is important that the last date for specifying standards is 3 years into the programme, giving vendors, and developers within the community, adequate time to develop and / or extend systems to support the specified schemas, as illustrated in figure 11.

Needless to say, if for any reason all the big vendors are not prepared to go along with this, we have to be prepared (and be seen to be prepared) to walk away from them and move to community led / funded development. This seems very unlikely however, as the UK academic trials market (and others who may find these standards useful and adopt them, e.g. via ECRIN) is not insubstantial.

Actions for the first 3 months

The first few months of any large programme like implementing data standards will be spent setting up infrastructure and finalising plans. It is suggested that if the programme goes ahead, and if money is made available for one or more of the projects identified in this document, the following will have to occur:

a) The construction of a board of stakeholder representatives (perhaps a reconstituted DIMS board) to oversee further developments.

b) The setting of detailed objectives in each of the main projects, and the establishment of suitable management and reporting arrangements for them.

c) Estimates of the costs involved in each of the main projects

d) Once provisional costing has been done, initial decisions to allocate budgets against projects and initial prioritisation and planning of effort.

e) The appointment of a small group of core staff, perhaps only one or two initially, needed to drive the programme through and do some of the necessary legwork. (Further secondment is unlikely to work beyond a few more months input).

f) There should be further detailed investigation and recording of current expertise, use and attitude towards data standards in the trials units.

g) There should be further detailed investigation and recording of current development projects and expertise and in the trials units, where those projects are relevant to data standards.

h) The initial invitations to membership of groups needed to take specific issues forward (e.g. construction of a trial administration data model) should be issued.
i) Involvement should also be sought in ECRIN, CDASH and the BRIDG project, (and any other groups thought to be useful, e.g. SNOMED users in the NHS) for members of ISWG and / or staff involved in the standards programme. If it is apparent that training in key concepts (e.g. HL7 RIM) is necessary to play a full role in these groups it should be provided.

j) Initial publicity materials about the project should be assembled and distributed, backed up by presentations at relevant meetings.

Actions for the first 12 months

Over and above the set up activities listed above, the following are suggestions for activity that could be included in the first 12 months of the programme, derived from but substantially more detailed than the various time frames described above. The actual targets will vary as a function of the resources made available.

(subject identifiers)

1. The NHS Number (or its Scottish / NI equivalents) should be recognised as the central patient identifier and must be collected, for UK residents, in all trials. Trials units should confirm they can sign up to this.

2. Clarification should be sought about the legal implications of returning the NHS number as part of the monthly accrual returns. If this is acceptable, and / or easily managed, UKCRN systems should be modified to deal with the imported NHS number, and the mechanics / costs / legality of identifying other demographic data using the NHS data explored.

3. Trials units should be canvassed as to their views on the other proposals regarding identifiers. If they are in agreement then those proposals (or a suitably amended version of them) should be actioned.

4. A time frame and plan should be drawn up for the introduction of encrypted identifier storage in all UKCRN units.

(pharmacovigilance)

5. For ISWG to liaise with the other UKCRN working groups and other relevant bodies to discuss the current use of MedDRA and the best ways to further promulgate knowledge about the system. To deliver a document summarising the current position.

6. For an overarching licence deal with MSSO to be negotiated centrally (though MedDRA is getting cheaper all the time and should be within the budget of most units, it would be much simpler to have a single agreement covering all academic units).

7. For existing standalone and web based coding tools to be investigated so as to better assess the best way of proceeding in terms of providing such tools generally. To cost and report back on options. To include using SMQs in the functionality.

(contact management)

8. For NIHR / ISWG to draw up a summary of current systems for storing / managing organisation and people data across trials units (especially those with systems dedicated to this), the NIHR (especially its identity management system), the NHS NACS and ODS systems, SUS systems and Trust / Network systems like EDGE.
9. For NIHR / ISWG to assess the viability of using the NIHR people / role / organisation model as the basis of a comprehensive contacts system, using e-GIF data structures, given the existing data and systems.

10. The legal implications of the sharing of information about professional staff in trials should be investigated and clarified.

11. The extent and types of information trials units need to hold about people and places from outside the UK, and thus normally outside NHS / NIHR systems, needs to be characterised. Alternative data capture mechanisms need to be identified for this information (and costed).

12. As a preliminary step units should make their data mappable to the e-GIF format.

13. One or more ‘development workshop’s will need to be held by all the major stakeholders (i.e. those listed in 8 above) to work out the most effective technical architecture for a loosely coupled (inter)national clinical trials contacts system.

(developing a metadata repository and a data item selection tool)

14. The ways in which the academic trials units / public sector can collaborate in the wider CDISC MDR project should be explored with UKCRC and other potential funders. That collaboration should begin as soon as possible and the results of the process disseminated widely. The details of many of proposals 15 – 19 below may alter depending on the nature and extent of collaboration that is possible with this project.

15. The roles and requirements for the metadata repository tools and their management should be carefully specified and circulated to relevant staff. Once the specification is finalised...

16. Any existing tools, developed within a trials unit, an academic unit, a pharma company or available commercially, that could fulfil or be further developed to fulfil the role of a trial design tool should be identified.

17. If a tool already exists arrangements need to be put in place to ensure long term central support (source repositories, documentation, testing and validation etc.), plus compensation for the originating unit(s). If one does not exist, or needs further development...

18. Those interested in the development work (either standalone and / or web based) should be invited to discuss architecture and design – e.g. during development workshops. Central involvement would be essential.

19. Estimates of costs should be made, funding found and a project established formally with timelines for delivery of the initial version.

(standardising on trial metadata description)

20. All those currently involved in this area in the UK should meet to exchange ideas, approaches and data. As much collaboration as possible should be sought.

21. An XML schema, based upon but extending the CDISC ODM schema, should be developed and published for consultation, to capture metadata about the trial’s data items, including skip and check logic.
22. An XML schema, based upon the Trial Model in BRIDG but also using the PCROM where that is seen as more relevant, designed to capture metadata about the trial’s schedules, visits and forms, should be developed and published for consultation.

23. Identification of how metadata extraction could occur for existing trials – hosted in different systems - should occur and solutions worked up and demonstrated. Using existing expertise in this area, across different trials units, will be a component of this process.

24. For new trials the metadata extraction mechanism should be built into the proposed trial design tool (see 17 above).

25. Work should continue within the relevant forums to influence the further development of metadata models and proposals, to ensure compatibility with metadata systems developed centrally.

(demographic and other non-clinical coding)

26. For candidate variables to be identified, initially concentrating on those common to large proportions of trials (e.g. identifiers, gender, date of birth, date of randomisation / registration, consent date(s) etc.), and for those variables to be matched against existing coding systems.

27. For the list to be presented to ISWG for comment and discussion of practical implementation issues and for a time frame to be drawn up for the introduction of these codes in new trials.

(developing a common data model for trial administration)

28. A summary should be drawn up of current systems for storing / managing trial administration data across trials units, the NIHR (especially the longitudinal research record) and Trust / Network systems like EDGE.

29. Existing systems should be compared with existing models, including the developing BRIDG model, to identify where there are gaps. ‘Borrowing where possible, inventing where necessary’, the model should be extended to fill the gaps.

30. Involvement should be established within the relevant forums (e.g. BRIDG) to influence the further development of formal models and proposals and to try and ensure compatibility with systems developed in the academic trials community.
**List of Interviewees**

Paul Amos  
ISB HaSC Domain Lead, Technical

James Batchelor  
Head of ICT at University of Southampton

Simon Bishop  
Head of GSK Clinical Trials Standards Group and Member of CDISC Technical Advisory Board

John Brazier  
Head of Data Services, MRC Clinical Trials Unit

Derek Coleman  
Epidemiology department, ICR

Jim Davies  
Professor of Software Engineering at the University of Oxford

Brendan Delaney  
Professor of Primary Care at The University of Birmingham

Ian Ford  
Professor Ian Ford, Robertson Centre for Biostatistics, University of Glasgow

Steve Harris  
Research Officer at Computing Laboratory, University of Oxford

Sharon Kean  
Director Information Systems, Robertson Centre for Biostatistics, University of Glasgow

Peter Knight  
NHS CFH Research Capability Programme Director

Paul Mason  
Head of IT, CRUK Clinical Trials Unit, University of Birmingham

Adel Tweedel  
Primary Care Clinical Sciences, University of Birmingham

Claire Snowdon  
Deputy Director, ICR CTSU

Rob Thwaites  
Chairman, ABPI Research Capability Programme Working Group

Ben Toth  
Founder, Health Perspectives

John Varlow  
ISB HaSC Domain Lead, Public Health and Statistics

Steve Walker  
NIHR - IS Programme Director

Adrian Willis  
Senior Programmer at University of Warwick
References

1. The NHS Connecting for Health’s Research Capability Programme’s website is found at http://www.connectingforhealth.nhs.uk/systemsandservices/research

2. The Office for Strategic Co-ordination of Health Research and takes an overview of research funding channelled through both the MRC and the NIHR. See http://www.nihr.ac.uk/about/Pages/about_oschr.aspx

3. The NHS-HE Connectivity project, or N3-Janet Gateway, has a website at http://www.nhs-he.org.uk/n3-janet-gateway.html

4. For more information on the use of SNOMED CT in the UK NHS see http://www.connectingforhealth.nhs.uk/systemsandservices/data/snomed

5. The NCIN is the National Cancer Intelligence Network and is part of the NCRI. It has a website at http://www.ncin.org.uk/

6. The Joint Initiative on SDO Global Health Informatics Standardization brings together CEN/TC 251 (the European technical committee on clinical data standards), ISO/TC 215 (the global technical committee on clinical data standards), HL7 and CDISC, It has a website at http://www.global-e-health-standards.org/

7. The IHTSDO (International Health Terminology Standards Development Organisation) is principally concerned with SNOMED CT. It has a website at http://www.ihtsdo.org/

8. For general information on the NHS number see http://www.connectingforhealth.nhs.uk/systemsandservices/nhsnumber/staff
   For more technical information see http://www.datadictionary.nhs.uk/data_dictionary/data_field_notes/n/nhs_number_de.asp

   For specific information regarding name and address data see http://www.govtalk.gov.uk/documents/GDS%20Catalogue%20Vol%202.pdf
   (Catalogue of name and address elements)
   http://www.govtalk.gov.uk/gdsc/html/noframes/imagemaps/PersonNameUML.htm
   (UML diagram of name). A variety of other technical specifications are available on the govtalk website

10. To view any of the six existing NHS data dictionary disease data sets see the website at http://www.datadictionary.nhs.uk/version2/web_site_content/pages/data_set_indices/data_sets_middle_pane.asp
11 To view the various datasets on cancers available from the Royal College of Pathologists go to the webpage at http://www.rcpath.org/index.asp?PageID=1091

12 Domains are a feature of the CDISC SDTM (Study Data Tabular Model). For a general introduction and a list of domains see http://en.wikipedia.org/wiki/SDTM. For more technical details see http://www.cdisc.org/models/sdtm/v1.1/index.html

13 The NCI’s Cancer Bioinformatics Grid, or CaBIG, has its web site at https://cabig.nci.nih.gov/ The page at https://cabig.nci.nih.gov/overview/ is recommended as an introduction

14 For a general introduction to SNOMED CT see http://en.wikipedia.org/wiki/SNOMED_CT
For a good general introduction to the way in which the system is organised see the user guide, available at http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_CT_Publications/SNOMED_CT_User_Guide_20080731.pdf

15 CDISC Clinical Data Acquisition Standards Harmonization; Details and documents available at http://www.cdisc.org/standards/cdash/index.html

16 Definitive descriptions of the CDISC MDR project are difficult to find on the web as it is relatively new. Probably the best starting point is http://asserowiki.com/index.php?title=Metadata_Repository on which most of the other relevant documents have been posted

17 The ePCRN has a web site at http://www.epcrn.bham.ac.uk/, though a better summary is found in Peterson K, Fontaine P, Speedie S. The Electronic Primary Care Research Network (ePCRN): A New Era in Practice-based Research, J Am Board Fam Med, 19 (1): 93, available at http://www.jabfm.org/cgi/content/full/19/1/93

18 The MRC data sharing (now data support) initiative has a website at http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?id=MRC003346

19 MedDRA - Medical Dictionary for Regulatory Activities; for overview and downloads http://www.meddramsso.com/MSSOWeb/index.htm (licence required for downloads)

20 For an introduction to several joint funded projects under the auspices of the UKCRC see http://www.ukcrc.org/infrastructure/expmed/expmedjoint.aspx

21 The ISO standard for metadata registries has a website at http://metadata-standards.org/ from which the standard documents can be downloaded

22 ICD-10, the International Classification of Diseases is a WHO sponsored system and has a web page at http://www.who.int/classifications/icd/en/ which includes an online version of the system.
In CDISC Meta data Repository: Enrichment & integration of CDISC Data Standards toward semantic interoperability, available as asserowiki.com/images/9/99/CDISC_MDR _PilotSpecification_v0.4.doc

The main CancerGrid website can be found at http://www.cancergrid.org/

The metadata registry can be downloaded from http://cancergrid.org/index.php?option=com_repository&Itemid=26&func=select&id=2

BRIDG is The Biomedical Research Integrated Domain Group, established by HL7, CaBIG and CDISC. For a summary see http://www.cdisc.org/standards/bridg.html For documents and downloads see http://www.bridgmodel.org/

A good browsable summary of the HL7 Reference Information Model can be found at http://www.miforum.net/distillate/rim/RIM0112_body.htm An access database version of the HL7 RIM is usually available from the sourceforge like site http://hl7projects.hl7.nscee.edu/projects/design-repos/

For details of the Longitudinal Research Record see the NIHR Information Systems Enterprise Architecture, available at www.nihr.ac.uk/files/pdfs/NIHR%20Enterprise%20Architecture%201.0.pdf

The IRAS (Integrated Research Application System) web site is available at https://www.myresearchproject.org.uk/. It is run by the NIHR

The CSP (Coordinated System for gaining NHS Permission) is run by the NIHR. Its web site can be found at http://www.ukcrn.org.uk/index/clinical/csp.html

OnIX is the NCRI Oncology Information Exchange, the website of which is at http://www.ncri-onix.org.uk/portal/#S1 For the home page of the NCRI informatics page see http://www.cancerinformatics.org.uk/


34 CDISC Operational Data Model Final Version 1.3; Details and documents available at http://www.cdisc.org/models/odm/v1.3/index.html


36 A good introduction to the HL7 Clinical Document Architecture can be had from http://hl7book.net/index.php?title=CDA which contains links to a list of CDA related resources

37 CDISC Study Data Tabulation Model; Details and documents available at http://www.cdisc.org/models/sdtm/v1.1/index.html

38 CDISC Analysis Data Model Version 2.0; Details and documents available at http://www.cdisc.org/models/adam/v2.0/index.html

39 For details on the mandated use of MedDRA in Europe see http://eudravigilance.emea.europa.eu/human/meddra01.asp


41 Edge is a trust based clinical research administration system, developed by the University of Southampton. For a press release see http://www.soton.ac.uk/mediacentre/news/2008/jun/08_118.shtml
### Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date Issued</th>
<th>Brief Summary of Change</th>
<th>Owner's Name/Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>07/12/2008</td>
<td>Initial draft version, sketched out document structure, some aspects of introduction, listed six priority areas for standardisation, more detailed proposals on increased use of identifiers</td>
<td>Steve Canham</td>
</tr>
<tr>
<td>0.2</td>
<td>18/12/2008</td>
<td>Completed initial preamble, added two more tasks to priority list, fleshed out proposals, especially on identifiers and shared contact database. Gave each proposal a better defined separate section.</td>
<td>Steve Canham, after input from Will Crocombe, Jim Charvill</td>
</tr>
<tr>
<td>0.3</td>
<td>01/01/2009</td>
<td>Applied formatting, front page and contents; Re-organised initial sections and added outline of general recommendations. Fleshed out remaining short term proposals</td>
<td>Steve Canham</td>
</tr>
<tr>
<td>0.4</td>
<td>19/01/2009</td>
<td>Added general recommendations section; also a few references to PCROM work. Tidied up figures. Added contents page.</td>
<td>Steve Canham, after ISWG meeting and presentations on PCROM, Edge</td>
</tr>
<tr>
<td>0.5</td>
<td>22/02/2009</td>
<td>First completed version! Re-organised and extensively rewrote four main sections on recommendations. Added section on time frames and vendors; re-organised Replaced section on analysis data sets with one on curation formats</td>
<td>Steve Canham, following in particular input from Brendan Delaney, Paul Mason, Adrian Willis, James Batchelor</td>
</tr>
<tr>
<td>0.6</td>
<td>15/03/2009</td>
<td>Comments from John Brazier incorporated Rewrote section on metadata repository to include references to CDIS MDR project and ISO 11179. Initial actions revised and extended slightly (by 2 further actions) to accommodate this.</td>
<td>Steve Canham, following in particular input from Simon Bishop</td>
</tr>
<tr>
<td>0.7</td>
<td>02/04/2009</td>
<td>Rewrote some parts of discussion of metadata repositories, trial data models, following input from Steve Harris, Jim Davies, Ben Toth Some graphics repasted as enhanced metadata files for greater clarity Minor revisions and corrections References added</td>
<td>Steve Canham</td>
</tr>
<tr>
<td>0.8</td>
<td>23/05/2009</td>
<td>Minor amendments</td>
<td>Steve Canham following DIMS board meeting, UKCRC CTU OG</td>
</tr>
</tbody>
</table>

This is an NIHR controlled document. On receipt of a new version, please destroy all previous versions.