UK CLINICAL RESEARCH COLLABORATION

igniting our potential

UKCRC R&D Advisory Group to Connecting for Health

Report of Research Simulations

Contents

Executive Summary 6
1. Introduction 9
2. The Simulations: Background & Objectives 10
3. Benefits - Potential Gains for UK Healthcare 12
4. The Simulations: Summary of Key Findings 15
   4.1 Surveillance 15
   4.2 Clinical Trials 16
   4.3 Prospective Tracking of a Known Cohort 17
   4.4 Observational Epidemiology 17
5. Recommendations 19
   5.1 Quick Wins 19
   5.2 Short Term Deliverables 19
   5.3 UK-Wide Strategy and Next Steps 20

Annex a. Full Reports of the Simulations 21
Annex b. Membership & TORS of Advisory Group 117
Annex c. Membership of the Simulation Subgroup 119

Acknowledgements 120

Abbreviations

ABHI	Association of the British Healthcare Industry
ABPI	Association of the British Pharmaceutical Industry
ADR	Adverse Drug Reaction
BIA	Bio Industries Association
BMA	British Medical Association
CHI	Community Health Index Number
CRDB	Care Records Development Board
GPRD	General Practice Research Database
GSL	General Sales List
HPA	Health Protection Agency
MEMO	Medicines Monitoring Unit, Dundee
MMR	Mumps Measles Rubella
NHS CfH	NHS Connecting for Health
NHS CRS	NHS Care Record Service
NPSA	National Patient Safety Agency
OTC	Over the Counter (medicines)
PCT	Primary Care Trust
PEM	Prescription Event Monitoring
POM	Prescription Only Medicines
SUS	Secondary Uses Service
THIN	The Health Improvement Network

Executive summary

The UK can significantly enhance its clinical research capability by using, strictly within the bounds of patient confidentiality, the electronic patient data that the UK's National Programmes for IT in the NHS have the potential to allow. This will have enormous benefits for all types of clinical, public health and health services research and for many aspects of patient care.

The UK Clinical Research Collaboration's (UKCRC) Research & Development Advisory Group to Connecting for Health therefore commissioned a series of simulations in October 2006 to provide the Department of Health Directorate of Research and Development, and NHS Connecting for Health (NHS CfH), with detailed specifications for a range of possible research applications. The objective was to:

- Inform future development of the NHS Care Records Service (NHS CRS)
- Highlight technical, regulatory and governance issues
- Inform plans for any further simulations and full pilots to test the capacity of the infrastructure, using real patient data with appropriate safeguards when this becomes feasible.

Four simulations were commissioned, based on a range of clinical research applications. These were: interventional clinical trials; surveillance; prospective tracking of an identified cohort; and observational epidemiological research. Detailed reports and key findings were presented to the Advisory Group in February 2007 and form part of this report.

The simulation leads worked as a team over this period and there was strong consensus in relation to both the high level and more detailed messages emerging from their work. They have identified a number of key data, regulatory and governance issues that need to be addressed for future development:

- Clinical services and research share the same mission of improving patient care and patient safety: **research is integral to patient benefit**
- Research makes a very important contribution to assessing the completeness and quality of data used for clinical care and health services
- Leadership is needed to create the sustainable and governance infrastructure required to exploit the research opportunities afforded by routine patient and other data
- Solutions should be addressed from a UK-wide perspective and build on the extensive experience with record linkage already in place
- Much of this research involves information about groups of patients rather than *individuals* and hence requires anonymised rather than identifiable data. However there will be occasions where data needs to be linkable (possibly by an 'honest broker') and comprehensive at the individual patient level in order to have maximum value and to allow quality and completeness to be validated
- Where data are required at individual patient level, such data access will need to be to pseudonymised data. Where identifiers need to retained, appropriate consent must be gained as part of enabling access to those data
- Existing UK strengths in the use of routine and other patient data for research will be significantly enhanced by the mandated use of a unique identifier (for example NHS number) in all

key NHS records and activities and by ensuring that new and existing data sets are person-based

- The data made available must cover the whole population, be up-to-date, and be retrospective over a number of years to give a rich historical picture of patients' health and care. They must also be accurate and based on high-quality input
- Further work to confirm the detailed requirements for data, which have been spelled out in each of the individual simulation reports, will need to be finalised. Much of the same data required for purposes such as safety monitoring and clinical trial research is of interest for public health and NHS management activity including monitoring service delivery. So there is a high degree of commonality in the data needs
- The breadth of data needed for the potential research applications explored in the simulations supports the concept of a data switchboard, with potential to link NHS Care Record data widely to other data sources. Thus future strategic developments should be based on this premise, rather than that of a single data warehouse
- A federated structure of data sources rather than a single data warehouse would also provide an effective infrastructure with optimal governance systems in place. This could be an honest broker with responsibility for removing identifiers, linkage of data and data quality checks

Tackling regulatory and governance issues successfully will be key to ensuring appropriate access and use of the data for research purposes.

- Ensuring patient confidentiality is critical. Data governance must be robust and at the same time capable of facilitating research
- Although for some research applications fully anonymised data will suffice, for many research applications pseudonymised data is required to enable linking of data sets or elimination of duplicate records. However, for other research purposes it will be important for patients to be contactable in an appropriate manner. Appropriate approaches to consent will need to be built into access mechanisms for information which might be capable of being linked to a specific patient
- In order to satisfy regulatory requirements for purposes such as pharmacovigilance and for clinical trials research, there are specific data quality and access requirements that need to be addressed

The dual role of the honest broker in ensuring patient data confidentiality and security as well as scientific integrity of data delivered to the research community will be key to engendering trust amongst patient, clinical professional and research communities.

The potential benefits for research will be lost unless these issues can be addressed.

It is critical that the needs of research be formally prioritised so that both individual healthcare and public health can reap the full benefits of this NHS resource. The recommendations are summarised below.

Key recommendations of the UKCRC simulations

Quick Wins

Recommendation 1: Mandate a common patient identifier

To enable linkage of sources of data at patient level a unique patient identifier will be required: use of the NHS Number should be mandated in all key NHS records and activities, including laboratory records.

Recommendation 2: Communicate the relevance of research to healthcare

There should be formal recognition that research is a core, not secondary, component of the development of the NHS Care Records Service as it benefits patients directly. Objectives, strategy and resources need to be committed or endorsed at the highest level of NHS Connecting for Health and reflected in its literature including website content.

Short Term Deliverables

Recommendation 3: Federate existing databases

A federated structure of data sources is required for research. A high-level strategy to support such an infrastructure needs to be developed together with a roadmap for its delivery. This strategy should ensure that the data made available cover the whole population, are up-to-date, person-based and of high quality, and extend back over a number of years to give a rich historical picture of a patients' health and care.

Recommendation 4: Improve data quality

Data quality is of paramount importance both in the clinical setting and for research. Data should be accurate (relying on high quality input) and based on a set of standards for recording and processing data. Ongoing processes will need to be developed to improve data completeness and quality which could involve development of incentives.

Recommendation 5: Initiate governance discussions

Tackling regulatory and governance issues successfully will be key to ensuring appropriate access and use of the data for research purposes. Data governance must be robust and at the same time capable of facilitating research.

Recommendation 6: Engage key stakeholders

It is essential to engage professional audiences who are key to implementation, particularly for the enhancement of data quality and improving data access. Patient safety is of importance to all audiences and should be at the forefront when communicating the value of research. A communication strategy regarding the joint benefits of using patient data for research and clinical care needs to be developed. The responsibility for this development and those in recommendation 5 above will be with the Care Record Development Board and may subsequently transfer to the National Information Governance Board upon its formation.

UK-Wide Strategy: Next Steps

In informing plans for next steps, the outcomes of the simulations suggest that more extensive data are required to enable research than those currently available through the Secondary Uses Service.

We recommend that an approach that relies on a federated system of databases should be based on a UK-wide strategy.

This will require:

- Initiation of pilots to link datasets, on the basis of existing successful examples within the UK;
- Definition of methods of access to the different sources housing the data. This should include access to detailed patient-level data from primary care, pathology services, disease registers and key private sector services;
- Future development which learns from, and build upon, existing skills, knowledge, databases and systems that have been developed in the UK over many years;
- Adoption of a UK-wide approach: not only are there examples of good practice beyond England that can be built upon, but the future development should ensure compatibility across the UK;
- An organisation capable of managing the specification and delivery of the required infrastructure and providing the linkage and definition of research support services.

1 Introduction

On 1 December 2005, the Chancellor announced a new commitment to develop the capability of the NHS National IT System to facilitate the recruitment of patients to clinical trials, and the gathering of data to support groundbreaking work on the health of the population and the effectiveness of health interventions. It was a strict criterion for this work that patient confidentiality would be protected throughout.

In this context, the UK Clinical Research Collaboration (UKCRC), the Department of Health and NHS Connecting for Health (NHS CfH) recognised the potential benefits for research of the developing NHS Care Records Service (NHS CRS)[1]. At the same time, reports from the Council for Science and Technology[2] and the Academy of Medical Sciences[3] highlighted the potential of NHS CRS to accelerate our understanding of health and disease by enhancing public health research on a national scale. This led to a constructive dialogue on turning this potential into a realisable proposition. As a result, the UKCRC and NHS CfH committed themselves to establishing a joint programme of work to move the dialogue beyond principles and generalities into relevant detailed work.

The UKCRC R&D Advisory Group to Connecting for Health is co-sponsored by the Department of Health Director of R&D (Sally Davies) and the Department of Health Director of IT Service Implementation (Richard Jeavons) and was established in July 2006. Its membership is drawn from the UKCRC and NHS CfH and includes cross-representation with the Care Record Development Board's Secondary Uses Working Group. Details of the membership of the Advisory Group are in Annex B.

The first stage in this programme of joint working has been to conduct a number of simulation exercises to investigate the feasibility of using the NHS CRS as a platform for healthcare research. The objectives and results of these simulations are explored in the following chapters.

In order to continue this important dialogue between the research community and NHS CfH, the Advisory Group has produced this report. We hope that its recommendations will lead to a programme of work for NHS CfH that will be beneficial to patients through the research that it enables.

As part of this exercise, we have sought to understand the similarities and differences in the approaches of the four UK nations. In order to implement a number of the recommendations from this report, NHS Connecting for Health, England will have to take steps to engage with national patient record programmes in the other UK nations.

Professor Ian Diamond,
Chair
The UK Clinical Research Collaboration Research & Development Advisory Group to Connecting for Health

1 Best Research for Best Health – A new national health research strategy (Department of Health, January 2006)
2 Better use of personal information: opportunities and risks (Council for Science and Technology, November 2005)
3 Personal data for public good: Using Health Information in Medical Research' (Academy of Medical Sciences, January 2006)

2 THE SIMULATIONS: Background & objectives

Modern health services cannot advance or be delivered for the benefit of patients without clinical, public health and health services research. This research is conducted at many levels. Every time a patient is provided with a treatment, an element of the unknown is present. New developments in managing patient data open up enormous opportunities for improving patient safety and enhancing the effectiveness of clinical treatments in the UK through research.

The opportunities lie in accessing the core information about patients and individuals that exists within various data systems and structures. These data can alert a doctor as to whether an *individual* patient displays characteristics that might render a particular medicine or mode of treatment dangerous. They can also be used to identify groups of patients that may be suitable to be invited to take part in a clinical trial, or that may be receiving the same treatment in order to determine its safety. Patient data that is currently routinely collected could be analysed, overnight if needed, to look for emerging patterns of disease or adverse reactions.

By establishing connections between databases such as those holding primary care patient records and cancer registry records, trends and associations can be explored which may have huge potential impact on patient outcomes and which can also support advances in biomedical research. In order to achieve such data linkage, we will need to address a range of capability and governance issues, which will arise in a similar way in a number of research applications.

To help clarify the potential for this new approach to the use of patient data, the UKCRC R&D Advisory Group to CfH commissioned four simulation exercises to:

- Inform future development of the NHS CRS and its UK equivalents
- Highlight technical, regulatory and governance issues
- Inform plans for any further simulations and full pilots to test the capacity of the infrastructure, using real patient data with appropriate safeguards when this becomes feasible

The four simulations commissioned were based on the following research applications:

Surveillance, including improved pharmacovigilance, post-marketing assessments, drug interaction and utilisation studies – led by Dr John Parkinson, Medicines and Healthcare Products Regulatory Agency and Steve Mott, DataPharm Ltd.

This simulation aimed to establish a set of requirements and a framework for implementation of an "ideal" surveillance system of patients' responses to medical interventions. The simulation team considered primarily Prescription Only Medicines (POM), including immunisation and over the counter (OTC) drugs as well as Devices. Additionally this simulation team sought to comment on surveillance of disease and general patient safety where there are commonalities between the requirements for data flows.

Interventional clinical trials including feasibility assessment, identification and recruitment of patients and remote electronic data capture – led by Rob Thwaites, GlaxoSmithKline Research & Development Ltd.

This simulation sought to investigate what is required from the NHS CRS to support world class clinical trials research. The team also conducted a protocol assessment using data currently available through the General Practice Research Database in order to identify issues and future needs from the perspective of clinical trials research. In particular, this team sought to highlight the key issues around data quality; data access and governance; and supporting systems.

Prospective tracking of an identified cohort of patients to flag up specified interventions or health outcomes – led by Andy Harris, UK Biobank

UK Biobank requires highly detailed data from many areas of NHS and other health related activity. This simulation team sought to investigate the extent to which the Secondary Uses Service is producing extracts that are relevant to UK Biobank's plans to:

- Study the underlying mechanisms of disease
- Identify markers of disease and disease progression

- Identify genetic, environmental and/or lifestyle factors predicting, pre-disposing or affecting disease onset or prognosis
- Support future studies based on UK Biobank that are of sufficient power to address new questions

The team also sought to understand how access to patient data depends upon implementation of governance arrangements as well as technical issues.

Observational epidemiological research based on retrospective analyses of data using case-control and cohort study design – led by Professor Carol Dezateux, and Professor Catherine Peckham of the Institute of Child Health, University College London, and Great Ormond Street Hospital for Children NHS Trust.

The aim of this simulation was to assess the extent to which the current arrangements proposed for the NHS Care Records Service and, in particular, its Secondary Uses Service, might support the conduct of high quality observational epidemiological research studies based on retrospective analysis of routine data.

This simulation used research case studies based on the outcomes and safety of assisted reproduction technologies as a model to address this issue, and to explore the capacity to:

- Link records of different individuals (for example, mother and child, offspring of the same birth mother)
- Combine records from different health providers (for example, private fertility clinics and NHS-delivered primary or secondary care maternity services) and from different health sectors (for example, primary care, pathology services, hospital episode systems and specific disease or treatment registers)
- Access data on potential confounding or effect modifying factors such as gestation, parity, and multiple births
- Access crucial legacy or historic data on earlier exposures

All four simulations aimed to identify and anticipate developments needed within the NHS CfH strategy to support their research application(s) and ensure a joined up approach across the UK. They were carried out over a period of four to five months from October 2006, with the four simulations coordinated and functioning as a team.

The key findings from these simulations are summarised later in this report and full details of the four simulations are to be found in Annex A.

3 Benefits:
Potential gains for UK healthcare

Introduction

This section highlights the potential gains for UK healthcare identified by the simulation teams.

Making the health service safer

Access to electronic healthcare records will enable early detection and therefore more timely response to adverse events including drug reactions. It will bring benefits to clinicians and patients by making the health service safer. It is estimated that 6.5 per cent of people admitted to hospital have experienced an adverse drug reaction (ADR). In 80 per cent of these cases, the ADR was the direct cause of admission. Two per cent of patients admitted to hospital with an ADR died.

An important use for electronic healthcare data would be to track the early patient use of new medicines (including vaccines) almost in real time. Both their effectiveness and any complications could be seen quickly enough to be acted upon. An example of where such a system would have had benefit is in dispelling the public confusion which followed the 1998 publication of a paper claiming a link between the childhood MMR vaccination and the onset of autism.

It will become possible for treatment to be varied according to real-time findings from the whole UK population, and for patients in hazardous situations to be treated appropriately. Inappropriate prescribing and treatment would become apparent, more readily and more rapidly.

Systems to enable earlier detection of, and more timely and effective response to, adverse reactions will improve the health and welfare of patients, the efficiency of the healthcare system and the economic viability of the pharmaceutical industry. In turn this enhances public confidence in the health care system. It can also facilitate the use of decision support systems which can reduce drug prescription errors, one avoidable factor in adverse drug reactions.

Reliable assessment of different causes of disease

Scientists have known for many years that our risks of developing different diseases are due to the complex interplay of factors such as lifestyle and environment, susceptibility (genes) and the play of chance (luck). However, despite this longstanding awareness, a clear picture of the combined effects of different factors on the risks of different diseases in different circumstances is yet to emerge.

Large studies with associated specimens stored (plasma, cells etc) in a range of settings, with prolonged and detailed follow-up of cause-specific morbidity and mortality, are needed to tease out the ways in which these three factors combine to cause disease. One such study is the UK Biobank, the setting for one of our simulation exercises. It aims to include 500,000 people from across the UK who are currently aged 40-69 and who are at risk over the next few decades of developing one or more of a wide range of important diseases including cancer, heart disease, stroke, diabetes, and dementia.

The NHS throughout the UK treats the single largest group of people anywhere in the world, and keeps detailed records on all of them from birth to death. Therefore, prolonged follow-up of participants through routine medical and other health-related records will allow the identification of comparatively large numbers of individuals who develop each of a wide range of disabling and life-threatening conditions. This access to and linkage of data will help researchers to understand the causes of diseases better, and to find new ways to prevent and treat many different conditions.

Identifying effective treatments more rapidly

The UK has a significant opportunity to increase clinical trial activity through the successful development of electronic patient records with research-focused national programmes for IT in place. Clinical trials are believed to be the most reliable method for establishing whether a treatment works and whether it is safe. Randomised double blind trials where the patient and their doctor are unaware what treatment is being given mean that the effects of any treatment can be assessed without influence from the prior beliefs of doctors, patients or any other interested party. They form the cornerstone of progress in healthcare worldwide. They benefit patients, the health service, and the UK economy.

While the UK has traditionally been a leading location for clinical trials research, the level of activity here in recent years has declined and newer, lower cost sites in other countries are now taking an increasing share of clinical trial activity. Access to comprehensive electronic patient level records (incorporating up to date and historical data from primary, secondary and tertiary healthcare settings, pharmacy and linked to laboratory results) would enable clinical trials to be conducted more efficiently, predictably and cost-effectively. This would make the UK a more competitive location for clinical trials, to the benefit of patients, the health service and the economy.

Longer-term access to comprehensive patient data would also improve long-term tracking of participants in clinical trials.

Answering public health concerns

Large-scale epidemiological studies have a distinguished tradition of contributing to the prevention and control of disease in populations, often before the exact biological mechanism has been identified. This is done by analysing variations in the disease and their associations with potential risk factors by time, place and person. Linkage of routine health service data, environmental, fiscal and educational data, and data on vital events such as births and deaths can enhance our understanding, and strengthen the scientific basis of strategies to maintain the health and wellbeing of the population and the prevent and control disease.

Although randomised controlled trials will continue to offer the highest level of evidence, they are not feasible or appropriate in all circumstances or for some hypotheses which they cannot be used to test. Using large linked datasets that go back over many years means that research can be carried out very rapidly without the need to wait to accumulate person-years of observation or events as in prospective studies. This is particularly helpful when dealing with uncommon outcomes or uncommon risk factors, where very large-scale evidence is needed, or with attempts to assess very long-term outcomes. In some such cases there could be an interval of up to a generation between exposure and outcome.

Linked data studies can also be useful to evaluate the effectiveness and safety of new technologies – including operation procedures, devices and diagnostic tests – when clinical trials may not be feasible. Retrospective observational epidemiological studies can also inform the response to unanticipated public health concerns about the safety of health service interventions (including medicines, immunisations, operative procedures, medical devices and transfusions) or environmental exposures (for example to radon, mobile phones or landfill sites) in a timely fashion.

International learning

A number of countries have established authoritative and planned linked systems which utilise a single NHS number or its equivalent in all settings. Here we look at developments around the world and within the UK.

A leading example is Denmark, where the citizen's number is the de facto NHS number. Here a powerful set of birth, hospital, death, twin, assisted reproduction, birth defect, prescribing, screening, immunisation, disease and device registers have been developed as an integrated system. They are linked in a framework with appropriate governance and have allowed a range of trials, patient and public safety studies and epidemiological research to be carried out. Similar systems exist in Norway, Sweden and Finland. Western Australia also has a very robust record linkage system and similar models exist within certain provinces of Canada and states or health care providers in the US.

Within the UK, Scotland has pioneered systems of linkage since the mid 1970s, when a far-sighted director of public health introduced the Community Health Index, the Scottish equivalent of the NHS number, and made it his mission to get this used in all NHS settings.

Hallmarks of the most successful systems include their use of a single unique identifier, their scale, and the close relation between research and clinical and public health services. They tend to include feedback between the researchers and the community that supplies and maintains the data, and invest effort in assuring data quality and completeness. This allows patients, the public, medical practitioners and researchers to respect and trust these systems as authoritative and reliable sources of information.

The UK experience base

The existing systems in the UK have already been used to good effect and we have a strong track record and tradition of excellence on which to build. For example, the links between vital statistics data and cancer registers are relatively well developed, allowing individuals to be traced. The NHS Numbers for Babies programme has been highly successful, ensuring that more than 98 per cent of babies are issued with NHS numbers shortly after birth.

Electronic medical record systems for general practitioners have been developed over many years and electronic primary care databases based upon these have been used to good effect to study adverse effects of drugs or immunisations, or the associations of disease. Primary care databases in the UK are recognised as being among the best in the world for research purposes.

Very large studies carried out in Scotland have allowed researchers to reliably identify the risk that a woman who has experienced a stillbirth or caesarean section will face in her subsequent pregnancy, as well as the risk to her infant.

In the North West of England, some relatively small but comprehensive and integrated databases have been successfully established. Each covers about a quarter of a million patients with the support of the local PCTs and Trusts, mainly to provide information to local medical professionals. These local health intelligence hubs (a combination of the integrated database and the capability to create and analyse the data) are beneficial for research, public health and, perhaps most significantly, for patients and those delivering healthcare in the local community.

Public health surveillance of HIV and other infections depends on record linkage and has been useful in developing policies, for example, by demonstrating the need for antenatal screening and its effectiveness. It also provides the only reliable way to monitor the changes in HIV infections occurring as a consequence of changes in the population, and in treatment and in drug resistance. But we could do more and the new electronic patient record developments create exciting opportunities to do just that.

4 THE SIMULATIONS: Key findings

Many branches of medicine use patient data derived from large samples of the population. We believe that there is now scope for research and clinical practice to take the use of patient data to a new level, with substantial benefits for patients, the public and health services. The opportunities lie in information systems and data structures that recognise connections between patients, and fully enable data linkage. This will lead to improved patient outcomes and advances in clinical, public health and health services research. However, it can only be achieved if key problems, including data access, quality, public and professional engagement with this new way of working are solved.

There are a large number of potential research applications that could benefit from, or extend the NHS Care Records Service. Many of these applications will depend on our ability to address a range of capability and governance issues, which will arise in a similar way in a number of potential research applications.

The proposals that follow do not call for a significant national increase in the gathering of medical data. Instead, they suggest integrating data that are already collected in new ways, to reveal patterns that have hitherto not been apparent.

All simulations identified common elements as will be apparent from scrutiny of the summaries below and the individual reports in Annex A. There was strong consensus that:

- Clinical services and research share the same mission of improving patient care and patient safety: **research is integral to patient benefit**
- Research makes a very important contribution to assessing the completeness and quality of data used for clinical care and health services
- Leadership is needed to create the sustainable infrastructure required to exploit the research opportunities afforded by routine patient and other data
- Solutions should be addressed from a UK-wide perspective and build on the extensive experience with record linkage already in place
- Although much of this research is directed at information about patient groups rather than *individuals* and hence requires anonymised rather than identifiable data, the data need to be linked and comprehensive at the individual patient level in order to have maximal value and to allow quality and completeness to be validated
- The breadth of data needed for the potential research applications explored in the simulations suggests a concept of a data switchboard with potential to link NHS Care Records data widely to other data sources, rather than as a single data warehouse
- A federated structure of data sources rather than a single warehouse would also provide an infrastructure with optimal governance systems
- Linkage of data within the Secondary Uses Service with primary care and pathology service data is a matter of priority

The key findings of the individual simulations are described here, and examined in more detail in the individual simulation reports (Annex A). Each simulation examined the suitability of the currently specified Secondary Uses Service (SUS) and the NHS Care Records Service (CRS) development to support their particular research application and identified where developments are required.

4.1 Surveillance (Pharmacovigilance)

The vision for an ideal surveillance system is of a nationwide “active” system for tracking patients’ responses to medical interventions (POMS, immunisations and OTC drugs as well as Devices) and of disease and other incidents requiring reporting.

The team concluded that the NHS CRS as a whole could support an advanced and indeed world-leading pharmacovigilance and disease surveillance system. Pivotal to this are the benefits delivered by an electronic healthcare records service based on the cradle to grave cover of the NHS. This provides added advantages over e-healthcare systems under development in other nations such as the USA. However mapping the requirements for such a surveillance system against the currently specified Secondary Uses Service (SUS) and the wider NHS CRS development led this simulation team to make a number of observations, including:

- The datasets currently available within the SUS

do not meet the necessary requirements

- The data from full clinical records including text are required. Omission of data from sealed envelopes may prompt recognition of adverse events and limit early ones, hence we recommend access to data in sealed envelopes emphasising that it is the event and not the individual that is of interest
- Extensive historical data are a requirement for much current drug safety research
- Data at the research level should be effectively anonymised but there needs to be a way back to the patient (either via their physician or honest broker) for the purpose of validation of the integrity of the data, or to follow up adverse events
- An NHS-wide standard incident reporting system is required for adverse events other than adverse drug reactions where an accepted path to the MHRA already exists. It should provide adequate coding to distinguish types of incidents so that the details flow to the correct authority
- Data streams from hospitals and day care cases need to be available at the level that is currently available from primary care
- Data streams from care delivered "on contract" to the NHS should be available as part of the full record

Existing pharmacovigilance activities in the UK are already some of the best in the world. However, there is room for improvement in these systems to provide an even more powerful system that would benefit patients, enhance the NHS information capabilities and lead to improvements in the way the pharmaceutical industry sector in the UK develops and makes available new medicines and treatments.

4.2 Clinical Trials

Within the range of activities involved in running a successful clinical trial in the future, there will be a need to access and process data from electronic records at a number of stages before, during and after the trial. The characteristics of data and systems needed for clinical trial research highlighted in the simulation include:

- Comprehensive data, covering as wide a population from the UK as possible to open up a broad subject recruitment pool and a representative population for trial planning and modelling
- Complete data across all care settings and services, up-to-date and legacy data to give a rich picture of patients' health and care. Data should be accurate (relying on high quality input). Data linkage at an individual patient level will also be needed, in order to build a complete picture
- Access to data must be governed strictly, though it is critical that researchers have access to patient-level data. Anonymised data will be suitable for many activities, including trial planning whilst through an appropriate, governed mechanism access to patient-identifiable information for recruitment to studies is required
- Supporting systems need to include: sets of standards for the recording, processing and transfer of data; a validated environment for the data; and specific processes and tools for clinical trials research
- A first-class capability for supplying data and associated services from a federation of data sources is needed, whether it is an enhanced SUS or a new development of infrastructure.

The UK has a long and well-respected heritage of generating and using electronic patient-level data for clinical research and it will be important to build upon this. Much of the data required for clinical trials research will not be available through SUS, and a federated system of linked databases should be adopted to build complete patient-level data.

An organisation with the capability to manage the design and delivery of the required infrastructure, and then to provide the ongoing linkage and research support services needs to be defined. In order for this organisation to successfully create and deliver the environment for clinical trials research alongside the concurrent development of other aspects of the National Programmes for IT, its direction, strategy and resourcing will need to be endorsed at the highest level of NHS CfH.

4.3 Prospective tracking of a known cohort

This simulation team concluded that in order for UK Biobank (which is a resource for prospective studies) to be able to provide the benefits of accuracy of assessment of, for example environmental exposures, the following need to be in place:

- Access to data will have to be at patient level data (identifiable) both coded and textual
- Clear and reusable/transferable rules must be defined between the provider and receiver of data which specify the use to which the data will be put, the security arrangements that will ensure the protection of the data (including confidentiality) and how these arrangements will be routinely audited
- Cohort Management – mechanisms are required to maintain and update the permissions for research projects to access demographic and patient record data
- In order to achieve its objectives UK Biobank requires access to the complete medical record of consented participants. This information will come from a range of sources but the NHS Care Record Service is expected to be the primary source. Information required includes that which is contained in the "sealed envelope" and the "sealed and locked envelope"

4.4 Observational epidemiology

The construction of retrospective observational epidemiological studies based on routine data sources requires access to data from a very wide range of electronic records, both within and without the health services. Evidence from other countries with strong research programmes based on routine data sources suggests that achieving this will require strong strategic leadership as well as resourced and governanced policies to create a long-term sustainable infrastructure.

The review and simulation undertaken confirmed the importance of taking steps to ensure that, in future, NHS data structures and systems include the following features or functions:

- Ability to combine data from different sources comprehensively at the individual level and family level using a unique patient identifier. This could be achieved relatively simply for data sources within the current NHS Care Record Service by building on the achievements of NHS Numbers for Babies. The use of the NHS number could be mandated as a unique identifier in all health care settings and encounters, starting with key maternity and birth datasets
- Ability to link data sources at the individual level together with access to individual patient data: although observational epidemiology is concerned with exposures and outcomes in groups of individuals rather than in individuals per se, this is essential to allow high quality and valid studies to be conducted and will require provision of appropriately governanced linkage and research support services
- NHS data sources based on people: to facilitate such linkage, all data sources should be person-based rather than episode-based
- The capability to link from anonymised data back to real data for checking and validation: while analyses for observational epidemiological research will be based on anonymised and pseudonymised data, an option of linking back is needed to confirm data quality and internal validity
- Inclusion of primary care, laboratory services and disease registers in the Secondary Uses Service
- Inclusion of data on private sector treatment and outcomes within the NHS system: mechanisms to achieve compliance with this should be explored
- Access to key legacy data as well as preservation of demographic histories: current arrangements for ensuring this are unclear and need to be addressed
- Ongoing processes to improve completeness and quality of data: incentives to ensure data are complete and of high quality are required, while recognising that research provides an important mechanism for evaluating and improving their quality
- New approaches to creating an authoritative suite of disease registers are needed with

which to capture key clinical diagnoses and outcomes of interest. UK disease registers are not comprehensive and are of variable quality, completeness and sustainability

- Development of the scientific and technical capacity for linking, curating and analysing large scale datasets

The simulation findings indicate that there are both immediate and longer-term gains to be made from better use of integrated patient data, and they highlight the technical, regulatory and governance challenges that currently exist. In the next section, these findings are developed into a set of recommendations.

5 Recommendations

Introduction

This section includes the key recommendations emerging from all four simulations. These recommendations are divided into '**Quick Wins**', '**Short-term Deliverables**' and '**Development of a UK-Wide Strategy**'.

Because much of the data required to realise the potential benefits will not be within the Secondary Uses Service as it is currently envisaged, these recommendations are aimed both at NHS CfH, and at a wider range of stakeholders including those who fund UK research. It is hoped that the chosen path for future infrastructure development will take on board the recommendations.

If acted upon, these recommendations could realise the vision of a coherent strategy for use of medical data and for other data (e.g. social care & environmental data) which is of potential use in improving health. This strategy will need to be sustainable over the long term. The culture we want to develop is one in which research, using the electronic patient record is seen as producing direct benefits for patients.

5.1 Recommendations: Quick Wins

In order to build a complete picture of each patient's health and care, data linkage at an individual patient level will be needed. Pulling information together from different sources for a patient will require a unique identifier for each patient, and investment in establishing and maintaining linkages across these sources.

Recommendation 1: Mandate a common patient identifier

To enable linkage of sources of data at patient level a unique patient identifier will be required: use of the NHS Number or its equivalent should be mandated in all key NHS records and activities, including laboratory records.

While clinical research is a core component of an effective health delivery system, the development of the NHS CRS currently does not specify research as a goal. This undermines the potential for developing the NHS CRS to capture the benefits of research. It is critical that the needs of research be formally prioritised so that both individual healthcare and public health can reap the full benefits of this NHS resource.

Recommendation 2: Communicate the relevance of research to healthcare

There should be formal recognition by NHS CfH that research is a core, not secondary, component of the development of the NHS Care Records Service as it benefits patients directly. Objectives, strategy and resources need to be committed or endorsed at the highest level of CfH and reflected in its literature including website content.

5.2 Recommendations: Short Term Deliverables

Access to patients and patient data

Data gathered within the system would include primary care, secondary and tertiary care, which could be clinical, day case or hospitalised care data, data from laboratory services, legacy data, demographic histories and improved disease registers. The research that can be performed with these data will be improved if it can capture as much legacy data as possible, and if its population coverage is as complete as possible.

There are already precedents for linking primary care and hospital data to deliver a fully integrated electronic healthcare record, for example pilots in the Wirral and Salford.

Recommendation 3: Federate existing databases

A federated structure of data sources is required for research. A high-level strategy to support such an infrastructure needs to be developed together with a roadmap for its delivery. This strategy should ensure that the data made available cover the whole population, are up-to-date, person-based and of high quality, and extend back over a number of years to give a rich historical picture of a patient's health and care.

These data would have to meet certain agreed standards and be ordered within a recognised data structure, using specific terminology and methodology. We also hope that connections will be developed between data in England and data held in other UK countries, so the standards applied should

be internationally recognised.

Recommendation 4: Improve data quality

Data quality is of paramount importance both in the clinical setting and for research. Data should be accurate (relying on high quality input) and based on a set of standards for recording and processing data. Ongoing processes will need to be developed to improve data completeness and quality, which could involve development of incentives.

For many research applications, pseudonymised data will suffice. But for some research purposes it will be important for patients to be contactable in an appropriate manner e.g. by their physicians. Appropriate consents will need to be built into access mechanisms for information which might be capable of being linked to a specific patient. Only researchers with appropriate levels of access would be able to approach patients directly to obtain information.

Recommendation 5: Initiate governance discussions

Tackling regulatory and governance issues successfully will be key to ensuring appropriate access and use of the data for research purposes. Data governance must be robust and at the same time capable of facilitating research.

In order to satisfy regulatory requirements for purposes such as pharmacovigilance and for clinical trials research, specific data quality and access requirements need to be addressed. Each of the simulations list data requirements for their research application in the individual simulation reports.

Recommendation 6: Engage key stakeholders

It is essential to engage professional audiences who are key to implementation particularly for the enhancement of data quality and improving data access. Patient safety is of importance to all audiences and should be at the forefront when communicating the value of research. A communication strategy regarding the joint benefits of using patient data for research and clinical care needs to be developed. The responsibility for this development and those in recommendation 5 above will be with the Care Record Development Board and may subsequently transfer to the National Information Governance Board upon its formation.

5.3 Recommendations: UK-Wide Strategy

In informing plans for next steps, the outcomes of the simulations indicate that:

- Most of the data required will not be available through the Secondary Uses Service as currently envisaged. Therefore an approach that relies on a federated system of databases should be adopted;
- A UK-wide strategy should be developed to make this happen,
- This will require:
  - Initiation of pilots to link datasets, on the basis of existing successful examples within the UK;
  - Definition of methods of access to the different sources housing the data. This should include access to detailed patient-level data from primary care, pathology services, disease registers and key private sector services;
  - Future development which learns from and builds upon, existing skills, knowledge, databases and systems that have been developed in the UK over many years;
  - Adoption of a UK-wide approach: not only are there examples of good practice beyond England that can be built upon, but the future development should ensure compatibility across the UK;
  - An organisation capable of managing the specification and delivery of the required infrastructure and providing the linkage and definition of research support services

Annex A - Reports of the Simulation Teams

Section 1: Surveillance Simulation 22

Section 2: Clinical Trials Simulation 43

Section 3: Prospective tracking of a known cohort 77

Section 4: Observational Epidemiology 94

Annex A Section 1

UKCRC Connecting for Health Research Simulation Report: Surveillance

Dr John Parkinson, MHRA (GPRD)
Mr Steve Mott, DataPharm Ltd

Contents

1 Introduction

2 Benefits

3 Current Situation and Vision
- *3.1 Pharmacovigilance of Drugs and Immunisations*
- *3.2 Pharmacy and General Sales List Medications*
- *3.3 Treatments Given Under Other Medical Paradigms*
- *3.5 Patient Safety*
- *3.6 Disease Surveillance*

4 Specific Data Needs

5 Key Issues Emerging

6 Summary

7 Appendix – Yellow Card

1. Introduction

The key aim of the simulation was to establish a set of requirements and a framework for implementation of an "ideal" surveillance system of patients' responses to medical interventions - the ideal vigilance system.

Primarily this covers Prescription Only Medicines including immunisations and reference to Pharmacy and General Sale Medicines (P & GSL categories) as well as Devices.

Additionally the report will include comment on surveillance of disease, other medical paradigms, and general patient safety where there are commonalities between the requirements for data flows.

Evaluation of these requirements would establish the suitability of the currently specified Secondary Uses Service (SUS) and of Connecting for Health (CfH) programme to support the research, regulatory and public health applications and identify where developments are required.

The requirements were elicited through input from:

- The UK Regulator, the MHRA, for medicines, herbals and devices
- The Pharmaceutical industry as represented by the ABPI (DataPharm), the BioIndustry Association (BIA) , the device organisation Association of the British Healthcare Industry (ABHI) and specific input from GlaxoSmithKline and AstraZeneca.
- Academics, existing database custodians and others involved in current Pharmacovigilance activities
- Consultation with other groups such as Health Protection Agency (HPA), National Patient Safety Agency (NPSA)
- Input from CfH/SUS
- Together with input from recently published reports; in particular the BMA report[1].

Prescription Medicines

The report covers all aspects of medicine surveillance and includes:

- early warnings of previously unrecognised potential problems (hypothesis generation);
- early warnings of changes in severity or incidence of known potential issues (hypothesis strengthening);
- access to person level, suitable anonymised, clinical and lifestyle related information in order to undertake hypothesis testing studies;
- ensuring medicines are given appropriately and that known problems that can arise are prevented (expert/decisions support systems);
- ensuring that new warnings, withdrawals etc are disseminated in the most rapid and appropriate manner consistent with a fully electronic NHS;

It will address, where relevant, the issue that no medicine is taken without some risk and that what is important is the balance of risk and benefit.

1 Reporting adverse reactions - BMA May 2006

As a consequence surveillance needs to take into consideration the effectiveness of medicines.

Existing pharmacovigilance activities in the UK and the datasets used are already some of the best in the world. However, there is complete acceptance from all parties that these systems can be improved to provide an even more powerful system that would benefit patients, the NHS and the way the pharmaceutical industry develops and makes available new medicines and treatments.

However, such changes and improvements are easier to document than to deliver in their entirety because the data volume is very high and in some cases the methodologies of very large dataset manipulation are not yet fully established. Additionally the skill sets for such work are in limited supply.

The systems described here would provide a mechanism for defining and implementing Risk Management Plans. These plans are now required at the time of submission for a licence.

Pharmacy and GSL medicines
Increasing amounts of these medicines are being taken and more powerful medicines are becoming available without prescription. There is therefore an increasing need to have surveillance in this area both to monitor their safety in widespread use and to improve the validity of research into POM where over the counter medicines can impact on both risk and benefit.

Other paradigms including herbal surveillance
There are some very limited forms of safety monitoring for herbal treatments. Herbals alone may cause serious adverse events but can often be associated with events when used in conjunction with prescribed medicines.

Devices
The use of devices is increasing as is their complexity. The area would benefit from improved capabilities of surveillance.

Disease and wider health surveillance
This report also describes some of the requirements of the Health Protection Agency in monitoring infections, radiation and chemical hazards and emergencies. Opportunities exist for widening and extending the current systems to enable near real time surveillance of these harmful agents.

2. Benefits

The primary benefit of pharmacovigilance is well established : patient safety with the appropriate level of risk and benefit. It is estimated that 6.5% of people admitted to hospital have experienced an adverse drug reaction and that in 80% the ADR was the direct cause of admission; 2% of patients admitted to hospital with an ADR died[2]. From a wider perspective the implementation of the recommendations of this simulation will help ameliorate the wider impact of Adverse Reactions, including clinical, economic and social factors.

Earlier detection and therefore more timely and effective response to adverse reactions is also an objective. Adverse events have a negative impact on the health and welfare of patients, on the efficiency of the healthcare system and on the economic viability of the pharmaceutical industry. The economic burden of a major drug withdrawal is significant on healthcare providers and the industry.

Improving the surveillance process offers the following benefits:

- A safer health care system for patients – due to earlier and easier identification of potential risk
- Improved information on drug risk and benefit
- Improved delivery of care – due to the ability for more real time feed back of potential risk
- Reduced number of patients who experience an unknown risk; a risk that has not been recognised cannot be equated against the potential benefit offered by the medicine

2 Reporting adverse drug reactions - BMA May 2006

- Ability to provide risk minimisation tools that ensure the knowledge about real world use of new medicines (including dose and indication changes) informs the understanding of best personalised use
- Maximises the ability of online decision support tools that can prevent inappropriate prescribing
- Improved public confidence in the healthcare they receive
- Reduced resource use in managing the outcomes of adverse drug reactions
- A unique data source for enabling the global pharmaceutical industry to more fully understand the role of their products in the real world by meeting their risk management requirements
- A surveillance system for medicines administered in hospital or day case settings. This offers a system that does not currently exist due to the low number of hospitals with electronic personal prescribing/ administration systems

In relation to other forms of surveillance these additional benefits are offered:

Other medical paradigms:

- The ability to include vigilance (risk and benefit) of other forms of interventional treatments by the inclusion of, or record-linkage of, data from healthcare records maintained in each of the other paradigms (homeopathy, acupuncture, herbals etc)
- The ability to assess and prevent drug and herbal interactions or other antagonistic or even synergistic effects of treatments taken at the same time.
- Ensuring that only effective and safe therapies are provided under the NHS or within any approved framework of care.
- Confidence that such treatments have the right balance of effectiveness and safety.

Medicines purchased from pharmacies or general sale:

- Medicines that are given a P or GSL licence are those that either have passed the test of time, as with aspirin, or are more recent introductions that have a suitable safety record. However, there remains a need to monitor such use and to record such use in a patient's healthcare record because this may impact positively or negatively on outcomes related to prescription only medicines. For example regular OTC aspirin use would positively affect the cardiovascular risk profile of patients taking prescribed cardiovascular medicines.

Device surveillance:

- Improved surveillance in this area would offer similar healthcare and economic benefits to those delivered for medicines.

Disease and wider health surveillance

- More rapid detection and response to outbreaks of infectious disease
- Collation and integration of data from a variety of sources to enable unprecedented coverage and comprehensive information for the evaluation of public health interventions, including immunisation programmes
- Improved surveillance of the determinants, spread and incidence of infectious disease
- Improved monitoring of radiation and chemical hazards and emergencies enabling improved research and response
- Improved ability to respond quickly to major incidents, including deliberate releases involving infection, radiation or chemicals

3. Current situation and vision

3.1 Pharmacovigilance of drugs and immunisations

Current situation

Spontaneous reporting

A passive voluntary spontaneous surveillance system - The Yellow Card Scheme (YC) (Appendix 1), run by the MHRA, has existed for many years in the UK. It relies on healthcare professionals voluntarily reporting their suspicions that a drug/herbal or drugs used in combination may be the cause of an unexpected event.

When compared to systems operating in other European countries and the USA, the YC has higher rates of reporting and works well. However, reporting rates are still considered low and many doctors have never sent in a YC. Evidence suggests only 10% of serious reactions are reported (serious reactions are defined as those which are fatal, life-threatening, disabling or incapacitating, resulting in hospitalisation, causing congenital abnormalities or which are medically significant). 2 to 4 % of non-serious reactions are reported.

Recently, patient reporting has been done using either a freephone number or a modified version of the YC. A unique advantage of patient reporting is that it includes reporting of events that could be associated with self-purchased medications and an accurate narrative description of the adverse event.

The YC allows any suspicion of an adverse reaction at any time during a medicines lifecycle to be reported. Healthcare professionals are asked to report all serious reactions. They are also asked to report all reactions (including non-serious) for medicines that are new or under detailed scrutiny (medicines denoted with a Black Triangle in prescribing information). Any reaction in children and any that are highly delayed from the original exposure should also be reported, regardless of seriousness.

Prescription Event Monitoring

A Prescription Event Monitoring (PEM) system, the Green Card (GC) system, which covers England only, has existed since 1980 to monitor adverse events for new drugs prescribed by GPs. This system is run by the Drug Safety Research Unit (DSRU) in Southampton. Green cards are sent to the cohorts of 12,000 – 15,000 (up to the first 40,000) users of some, but not all, new medicines. This is based upon data provided by the Prescription Pricing Division, which is part of the NHS Business Authority (NHS-BSA). Questionnaires are then sent to the GPs who prescribed the drug on which they are asked to record all events that occurred since the first prescription of the drug. This might include referral to hospital, hospital admission or death. For important events, follow-up forms are sent to GPs to obtain more information. The DSRU has completed over 100 PEM studies.

Both YC and GC are essentially systems designed to detect unexpected effects. In research terminology they are hypothesis-generating systems.

Databases for drug safety research

The hypotheses generated by YC, GC or other reports of cases have to be tested using well accepted but statistically challenging methodologies on person-level, anonymised healthcare data. A number of UK databases enable such research:

- The General Practice Research Database (GPRD) is a not-for-profit database owned by the Secretary of State for Health and managed by the MHRA[3]. It consists of essentially anonymous GP-recorded patient records from just over 5 % of all UK patients. The data is anonymous to the researcher and is, both coded and textual. The data is sourced from the VISION GP Clinical system, but importantly the accuracy of the data can be validated because there is a way back to the original data with the key held only in the clinical domain.
- Tayside, MEMO - this is an academic regional system that takes data streams from many sources and links them using an anonymised form of the Scottish NHS Number. It covers 0.7% of UK patients and 13% of Scottish patients. MEMO enables validation and has been able to undertake studies that include medicines prescribed in hospital care. MEMO has similar patient consent arrangements to GPRD.
- The Health Improvement Network (THIN) is a commercial database that is a smaller and less validated version of GPRD;
- QResearch is a large database using fully anonymised data from the EMIS GP clinical system; data is available to academics, the NHS and the Department of Health and its arms-length bodies.
- IMS Disease analyser is a commercial database taking data from the Torex GP system

3 GPRD is managed by the MHRA but under arrangements whereby the GPRD operation is kept as a distinct operating entity from the regulatory function of the agency.

Decision Support Systems
Existing clinical systems that enable prescribing or dispensing contain decision support tools that have in the main concentrated on preventing the prescribing or administration of medicines that have inappropriate interactions. Some also enable specific warnings to be flagged for certain types of prescribing. Such systems currently work at a local level and could be vastly improved and extended via a fully integrated healthcare record system.

Vision

An active surveillance system across all parts of the NHS and all professions prescribing, dispensing or administering medications - Primary, secondary and tertiary care (including day cases), pharmacy care and drop in centres.

Such an active surveillance system will enable all parts of the surveillance/pharmacovigilance process and includes:

- Improving the system for capturing/reporting suspicions of adverse reactions to YC.
- Enabling access to patient's full electronic record (coded, textual, primary, secondary, tertiary and pharmacy care) for signal generation as well as providing data for hypothesis testing research.
- Embedding specific decision support systems within all clinical systems that enable and ensure safe prescribing, dispensing or administration of medicines.
- Extending the reach of the system to include, in particular, herbal treatments in a patient's record but where possible all treatments including some details of medicines self administered.

Details of each part of such a system are:

A. Real time suspicion reporting - e-Yellow Card (e-Yc) - hypothesis generation
All clinical systems should include the ability for real time electronic reporting of possible adverse events by prompting all healthcare professionals to report events they felt were suspicious of being associated with a drug or a number of drugs. Such a system could use prompts on healthcare professionals' screens activated to varying degrees based on whether the drug was new to the market or was the subject of a particular concern. Should the health professional decide to report a suspicion, the e-Yellow Card should to a large degree auto-populate.

Patient reporting of adverse events could be enabled via each patient's Health space (https://www.healthspace.nhs.uk/) that could have a yellow card semi-automated reporting system.

The e-YCs should be directed to the MHRA within the current UK and European regulatory framework.

B. Fully linked or linkable electronic records from primary, secondary, tertiary and other care (dentistry, pharmacy, family planning, travel clinics, walk in centres) for both hypothesis generation (HG) and testing (HT).
The availability of a complete "anonymous to the researcher" longitudinal personal health record for the whole or close to the whole population providing the following details would add considerably to the way pharmacovigilance could be undertaken. Flows of data from the following would be required as stated in each category:

- GP record - patient demographics and lifestyle; prescribing, symptoms and diagnosis, laboratory and other tests, referrals, other treatments. Equivalent from dentistry, family planning, travel clinics, walk-in centres, catch-up immunisation campaigns as well as data from independent prescribers (nurses, pharmacists). Required for HG and HT
- Hospital clinic record - prescribing, symptoms and diagnosis, laboratory and other tests, referrals. Required for HT and HG
- Hospital Day case – prescribing/dispensing/ administration and other relevant clinical information. Required for HT and HG
- Hospital in-patient – prescribing/dispensing/ administration, symptoms and diagnosis, laboratory and other tests, referrals, other treatments and procedures. Required for HG and HT.
- Maternity care - Outcomes of pregnancy with mother-baby links. Required for specific research connected with drugs in pregnancy.
- Retail Pharmacy – dispensed record. Adds validity to research otherwise conducted on the prescribed record. The availability of the dispensed record would ensure that collection of prescribing data from all who prescribe would not necessarily be required. Beneficial for both HG and HT.
- Death - linkage to centralised records. Beneficial as essentially complete record as compared to that available from General Practice.

The data flows for signal detection (HG) can be far simpler than those for the testing. However, if delivery of all data is enabled for testing then the signal detection can take place on a sub-set of the data.

Where data are available for hypothesis generation, advanced data mining techniques would be needed; these are already being used to some degree by the MHRA, FDA and a number of pharmaceutical companies. It should be noted that the current capacity and methodology to move, store and analyse the significant volumes of data that will be generated is a limitation, though this will be eased as technology and methodology progress.

In making the same data available for HT this would extend the existing research currently undertaken with samples of UK data (GPRD, Q Research, MEMO) by improving the validity of the data, extending the types of data available and in some cases enabling studies that require access to very large volumes of data. It must however be remembered that such research is done in a retrospective manner meaning that historical data is required. Current drug safety studies in GPRD and other datasets are using data from as far back as 1990.

Also of note is the fact that manipulation of the data to assemble it for research and the statistical analysis of such data can only be handled by certain high-end research analysis tools rather than readily available data-management, database or spread sheet tools whose data limits and capabilities are highly inadequate.

It should be noted that great care has to be exercised in allowing both hypothesis generation and testing to take place in the same dataset. There are issues that need to be managed by adequate scientific governance of all activities to ensure that both activities are acceptable at the same time.

C. Improved Prescribing Decision Support system
Reduction of known potential adverse events can be effected by more stringent application of mechanisms that prevent error. In particular the

prevention of:

- the taking of combinations of medicines or medicine-herbal combinations that lead to adverse reactions, diminished effectiveness, or even excess effectiveness that in certain circumstances can be harmful;
- the taking of inappropriate medicines because of allergies or contra-indications;
- the use of inappropriate doses particularly for children, the elderly, those on many medications and others where dose adjustments are required based upon measured/ known factors.

Such systems are already incorporated into clinical systems but in some cases they have been switched off because they cannot work fast enough in limited time consultations. New more powerful web enabled solutions should ensure they cannot be switched off and can additionally provide the benefit of off-line analysis and feed back done on the server.

The systems/services should also provide a means of communicating with the healthcare professionals, incorporating services such as warnings and "Dear Healthcare Professional" letters, dosage changes, recall and withdrawal of medicines etc. It would add to and improve the existing drug interaction checkers that have been incorporated into systems for many years.

Further refinements could include warnings to be triggered for medicines that required a particular laboratory test to be conducted prior to treatment or regularly during treatment, if the conditions were not met. A second example could be reminders to ensure that certain drugs were never used in pregnant women.

Parts A and B of the Vision will generate large volumes of data and require access to the full longitudinal electronic record for all patients (estimated to be five terabytes per year of data based upon current flows).

In addition to capturing and generating information a mechanism must be put in place to manage, transform and deliver data to researchers and to provide the capacity to centrally analyse the data in an approved manner with acceptable governance. We propose such a centre of excellence, the National Pharmacovigilance Data Centre, which will concentrate the necessary expertise to deliver the benefits described above. The centre would be responsible for all aspects of pharmacovigilance not least of which is communicating the benefits of the system throughout the global research community.

The alternative to this, where such work is handled by SUS, is hard to envisage because there would need to be massive transference of existing skilled people and specific high end knowledge rather than building on existing resources. These resources are already at the forefront of existing pharmacovigilance services.

[An alternative but longer term model of data available to Pharmacovigilance could be based upon GRID technology so making the requirement to move large volumes of data less of an issue.]

3.2 Pharmacy and General Sales List medications

Current systems

Currently surveillance of issues related to these medicines relies on reporting by doctors and other healthcare professionals who become involved with potential issues or by patients directly into the Yellow Card Scheme.

Vision

The system proposed in section 3.1 will automatically improve reporting related to these medications. Additionally for medications available under a Pharmacy (P) licence a personalised P dispensing record could be made available via the ETP system to a central system thus further enabling research in this area.

3.3 Treatments given under other medical paradigms

Current systems

There are no specific/central systems used to monitor the safety of many, if not all other, paradigms.

Vision

There is the potential to have practitioners of other medical paradigms store patient data generated in the course of their consultations and treatments electronically into a specified database. As with GPs this data could then be available via central means for the purpose of conducting vigilance related to each medical paradigm. Importantly where herbs are being used this information could be merged with prescribing/ dispensing data for medicines (POM, P, GSL) so enabling vigilance related to herbal-medicine interactions as well as provide a warning system when known conflicts occur.

3.4 Devices safety

Current situation

The Devices Technology and Safety Division at the MHRA have two online reporting systems for device adverse incident reports. For NHS organisations, healthcare professionals, patients and the general public there is a user reporting system, introduced in 2001. Over 60 % of user reports are received online. For manufacturers there is a more sophisticated online system with a password protected workspace allowing initial, follow-up, final and trend reports to be submitted in compliance with the Medical Device Directives. These reports are automatically imported into MHRA's Adverse Incident Tracking System (AITS) database for medical devices.

Reports can also come in via non electronic means and these get manually entered into the AITS database at the MHRA. A report requires information on the Device, the Event, the reporter, and where appropriate the Patient.

Most NHS organisations use an internal 'Incident Reporting Form' that covers all types of incidents including those associated with devices. Most of these reports currently start on paper and are then entered into an electronic Local Risk Management System (LRMS). There are about seven computerised systems that have been adopted within the NHS. Sometimes healthcare professionals report directly to MHRA. Other times Risk Managers or Medical Device Liaison Officers review incidents on their LRMS systems and report those that they believe are of interest to the MHRA. There is still considerable variability across NHS Organisations concerning (knowledge of MHRA device) reporting and decisions about what incidents they think we are interested in. This is despite regular communications from the MHRA about adverse incident reporting.

The NPSA has a system that takes data from all systems on a batch basis into a centralised system. Incidents are not particularly well classified to enable automated streams of data to agencies such as the MHRA who have the responsibility for device safety. For instance only 10 % of device related incidents contain details about the make and model of the device.

The coding of devices and use of the nomenclature used to report incidents are not yet universally used. Within CfH the current DM+D uses NHS Purchasing and Supply Agency (PASA) codes to describe the prescribable devices on the system. However, with considerable encouragement from MHRA, CfH are now seriously considering the Global Medical Device Nomenclature for this purpose.

Vision

A system within the NHS that ensures that all incidents are electronically documented using agreed standard coding at as close as possible to the incident and then made available to a central system. The coding will then enable each type of incident to be transmitted to the organisation responsible for that event be that NPSA, MHRA, Department of Health Estates and Facilities, HSE etc. This system should allow updates to be reported and two way exchange between the reporting body and the relevant investigating orgainsation in order to maximize learning opportunities.

GP systems should allow reporting of device incidents using a methodology similar to the electronic card, but utilising a medical device adverse incident dataset.

The use of patient reporting using an automated reporting form via their Health Space or generally via an easily available web portal would be welcomed.

The use of Radio Frequency Identification (RFID) with automated links to electronic health records where appropriate is to be encouraged.

For devices that are implanted or are integral to or with the body in any way, the information associated with that implantation (product, batch nos, etc) should become part of the electronic record available for research. An example of a current need is the requirement for an assessment related to genotoxicity of hip replacements. Such studies have the potential to be conducted in the same manner as with drugs - case control or cohort studies using the full patient's anonymised record.

Nomenclatures for medical device adverse incident communications

- The DM+D should include all devices and should be based on the Global Medical Device Nomenclature (GMDN)
- The medical device related adverse event nomenclature should be based on international standards (ISO TS 19218).
- Snomed-CT already seems to be the nomenclature of choice for describing the effects of the incident upon the patient.
- A standard should be developed for medical device related reported events which should be based upon European medical device adverse incident report datasets.
- A regulatory outcomes/actions nomenclature should be developed in consultation with MHRA.

Data relevant to each type of incident should be transmitted in essentially real time to the organisation responsible for safety and regulation.

3.5 Patient safety related to prescribing and administration errors as well as general aspects of safety

Current systems

A national incident reporting system exists. In July 2005, the National Patient Safety Association (NPSA) Board identified some broad themes which are particularly important for future NPSA work, although all ideas to improve patient safety will be considered:

Some areas highlighted from our data on reported incidents and other evidence:

- Older people
- Womens' health (including obstetrics & gynaecology)
- Children and young people
- Orthopaedics
- (Cross-cutting) Communication problems (including issues of language, culture, literacy)
- (Cross-cutting) Delay in diagnosis or access to care

Areas where reporting rates to National Reporting and Learning System (NRLS) are currently low:

- Social, residential and nursing home care
- Primary care
- Patients with long term conditions
- Patients being treated in their own homes
- Issues of race and patient safety
- Issues of patient safety and contestability, plurality (new providers including independent sector) and choice

The NPSA has as its core objective the collection and analysis of patient safety data to inform rapid patient safety learning, priority setting and coordinated activity across the NHS.

The NPSA should work in partnership with agencies and activities that gather different sources of data, such as complaints, claims and coroners' reports, as part of its National Patient Safety Observatory to ensure that all deaths and serious harm associated with adverse events are identified.

Key relationships should include the NHS Litigation Authority and the confidential inquiries make it

more effective in this respect, including simplifying and encouraging reporting as well as including a new category of analysing risk-prone situations and anticipating adverse events.

Vision

A system within the NHS that ensures that all incidents are electronically documented using agreed standard coding as soon after the incident as possible and then made available to a central system. The reporting of adverse events should be simplified. NHS organisations should be required to develop local strategies to encourage reporting. Near misses and a new category of 'adverse events that could happen' should also be reported. The NRLS should be redesigned to make it easier for clinical staff to report on a confidential basis without fear of retribution.

Simplifying reporting is important if we want to engage with busy frontline staff. The experience of other high-risk industries which have well established reporting systems highlights the importance of all staff having immediate access to easy mechanisms for simple reporting such as a computer or paper reporting box. To promote more rapid and effective learning, reports should be confidential but not anonymous. In particular, rules on confidentiality of data should not block the identification of very serious adverse events, recognising the need to ensure that reporters are free from retribution. The opportunity should be taken to ensure that local reporting mechanisms cover both the NPSA reporting and learning system and the adverse incident reports required by the MHRA.

3.6 Disease Surveillance

Current systems

The Health Protection Agency undertakes surveillance of infectious diseases and other environmental threats to health through a range of local, regional and national surveillance systems. The purpose of these surveillance systems includes acute alerting and response to health protection threats, the longer term monitoring of trends and distribution in disease, hazards and exposures, and elucidation of the determinants of disease epidemiology and the natural history of diseases due to infections and other environmental threats to health. A key requirement of health protection surveillance systems, and one that is almost unique amongst disease surveillance systems, is the need for real time or near real time, capture and analysis of data in order to detect and inform the response to outbreaks and other emerging infectious disease problems.

The Health Protection Agency draws on data from a wide variety of sources within the NHS, and beyond, as well as using data derived from its own frontline units and laboratories. Among the many data sources and systems used to monitor the wide range of infectious disease and other environmental threats to health there are a number of core systems that underpin much of the Health Protection Agency's surveillance activity, these include:

- Voluntary confidential reports from microbiology laboratories (and some other pathology departments)
- Notifications of infectious disease made under the provisions of the 1984 Public Health Act and 1988 Public Health Regulations
- Pseudonymised clinical reports of HIV/AIDS
- Data from Primary Care reporting networks including the Royal College of General Practitioners network and the QRESEARCH and QFLU networks
- Outpatient return data, in particular from genitourinary medicine clinics (KC60)
- Vaccination uptake information, derived from child health and other relevant information systems
- Data on hospital patients including case reports of surgical site infections and other healthcare associated infections, and hospital episode statistics
- Mortality data from the Office for National Statistics

The amount of person identifying data, demographic information, and other information relating to clinical status or risk factors varies between these surveillance systems, but, as noted above, the common feature of many of these surveillance

flows is that the data are collated and analysed on as near a real time basis as is possible. Many of the surveillance systems are based on voluntary reporting by clinicians and pathologists (notifications are unusual in that reporting is mandated in law), and for some data flows, in particular laboratory reports that constitute the mainstay of surveillance for many conditions, data flows are supported by ad hoc electronic reporting systems that could be compromised by the implementation of National Programme for IT Systems unless those systems are explicitly specified to deliver the required data.

The scope of data capture for surveillance by the Health Protection Agency is wide. For example, for notifications of infectious disease it potentially encompasses all clinicians, while laboratory reports are collected from all microbiology laboratories, and vaccination coverage data are collated from all child health systems.

Vision

There is the potential of CfH/SUS to deliver a nationwide system that moves the necessary data (disease, microbiology etc) from all parts of the NHS to the Expert Surveillance Centres in real time and that can also support longitudinal monitoring of individuals to enable surveillance of the long term effects of infectious disease and other environmental threats to health. It will be important that the process of implementation of this system does not disrupt or compromise existing systems, and that the fully implemented system inter-operates with complementary, cost effective and validated schemes.

4. Specific data needs

In this section the data needs for five surveillance scenarios will be covered.

- Prescription Only Medicines: Primary care prescribing Active surveillance that includes e-Yellow card and data for hypothesis testing
- Prescription Only Medicines: Secondary, tertiary and day care prescribing Active surveillance that includes e-Yellow card and data for hypothesis testing
- Advanced Prescribing decision support
- Other paradigms
- Disease and wider health surveillance

4.1 Prescription Only Medicines : Primary care prescribing Active surveillance that includes e-Yellow card and data for hypothesis testing

A) Outline Data Needs		
A.1	Real world-Clinical	
A.1.1	As recorded in existing systems	YES
A.1.2	Interventional, recorded in existing systems,	
A.2	Special screen-clinical	YES, for e-Yellow cards
A.2.1	in existing system recording	Possibly into patient record
A.2.2	external portal recording	An e-Yellow card portal
A.3	Patient/ non clinical recording	YES for patient ADR reporting
A.3.1	Into existing system	Via Health Space
A.3.2	Into new system	Into e-Yellow card portal
A.4	Other- state	

B) Sources of Data		
B.1	Primary	YES
B.2	Secondary	YES
B.3	Tertiary	YES
B.4	Pharmacy	YES dispensed data and pharmacist prescribing
B.5	Patient	YES, self reported ADRs.
B.6	Others Healthcare -state	Dentists, Pharmacists, Nurses, Private health centres who prescribe and those who administer POMs
B.7	Others – Social care	No
B8	Death	YES
B9	Census	YES, Socio-economic data
B10	Existing registries	Probably not
B11	Others- state	

C) Temporal needs of the data
Retrospective data with a record going back as far as is possible- minimum 2/3 years, preferable 10 years, ideally more. Data needs to be current to weeks. Prospective data collections of specific data requirements should be enabled via special screens.

D) Size/population of dataset required
Drug safety is a UK wide matter and current data streams come from all regions of the UK. An English only system providing access to data from 50+ million people will suffice many needs but where there are regional variations in prescribing, NICE/SMC for instance, or in disease prevalence (cardiovascular- Scotland) data from other parts of the UK will be required.

Active surveillance on 100% of England would be ideal but is not a necessary prerequisite. An acceptable level would be anything above 60-65%, provided this had a good geographic spread. At 60-65% or above it would ensure that for new products with either slow take up or rapid take up that the system was adequate to deliver the necessary amounts of data.

E) Source data verification
The ability to validate coding is an absolute where such has not been done before. This can be done on a sample basis and is the reason for the pseudonymous nature of the required data.

F) Level of identification required
Pseudonymised - using an identifier that enables linkage back but where the data is effectively anonymised for the researcher - i.e. the key is held in the clinical domain or by an Honest Broker/ Trusted Third Party.

A level of geographic locality is required - region of country would be acceptable.

G) Dictionaries
Access to all dictionaries in current use within the clinical systems and mapping to older dictionaries that were used prior

Snomed CT (medical codes)
DM+D (Drugs, devices etc)

Mappings to:
- ICD 9/10 (medical codes- hospitals)
- READ codes (medical codes- GP)
- OXMIS codes (Pre READ system)
- MedRA (regulatory medical, adverse drug reaction codes)
- Existing drug code systems – British National Formulary (BNF), Anatomic Therapeutic
- Classification Codes (ATC)

H) Timings /data delivery/governance/pre-study
Active surveillance of drug use does not require 24/7 data delivery. Data delivered within a 7-10 day period would be ideal but even a 30 day window would be acceptable. However the volumes of data generated by such a system might suggest that a continual feed to a pharmacovigilance server system might be ideal.

Certain governance issues can be handled as essentially one-off approvals as is the case now (PIAG, Ethics) whilst scientific approval needs to be gained for each study. By way of example the MHRA use the input of an independently appointed group - ISAC (Independent Scientific Advisory Committee) that review and approve all protocols for research on both Yellow card and GPRD data.

4.2 Prescription Only Medicines – Secondary, tertiary and day care prescribing Active surveillance that includes e-Yellow card and data for hypothesis testing

This has similar needs to that for primary care prescribing except it should be noted that in many circumstances (in this type of care) there is the potential for three records: prescribing, dispensing and administration. The administration record is the ideal data source for pharmacovigilance. However each type of medicines giving needs to be worked through as there will be some cases where there will only be a prescribing and/or a dispensing record.

It should be noted that what is required for “best” research on drugs used in any part of care is a fully integrated research record. This could be achieved at the research level if such were not available within each patient’s clinical record.

Other details of the data needs will not be repeated- see Primary care section

A) Outline data needs		
A.1	Real world-Clinical	
A.1.1	As recorded in existing systems	YES
A.1.2	Interventional, recorded in existing systems,	
A.2	Special screen-clinical	YES, for e-Yellow cards
A.2.1	in existing system recording	Possibly into patient record
A.2.2	external portal recording	An e-Yellow card portal
A.3	Patient/ non clinical recording	YES for patient ADR reporting
A.3.1	Into existing system	Via Health Space
A.3.2	Into new system	Into e-Yellow card portal
A.4	Other- state	

B) Sources of data		
B.1	Primary	YES
B.2	Secondary	YES
B.3	Tertiary	YES
B.4	Pharmacy	YES (dispensed data and pharmacist prescribing)
B.5	Patient	YES, self reported ADRs.
B.6	Others Healthcare -state	Dentists, Pharmacists, Nurses, Private health centres who prescribe and those who administer POMs
B.7	Others – Social care	No
B8	Death	YES
B9	Census	YES, Socio-economic data
B10	Existing registries	Probably not
B11	Others- state	

C) Temporal needs of the data
Retrospective data with a record going back as far as is possible, minimum 2/3 years, preferably 10 years, ideally more. Data needs to be current to weeks. Prospective data collections of specific data requirements should be enabled via special screens.

D) Size/population of dataset required
Drug safety is a UK wide matter and current data streams come from all regions of the UK. An English only system providing access to data from 50+ million people would suffice all needs provided there were no regional variations in prescribing, NICE/SMC for instance, or in disease prevalence (cardiovascular- Scotland).

Active surveillance on 100% of England would be ideal but is not a necessary prerequisite. An acceptable level would be anything above 60%, provided this had a good geographic spread. At 60% or above it would ensure that for new products with either slow uptake or rapid uptake that the system was adequate.

E) Source data verification
The ability to validate coding is an absolute where such has not been done before. This can be done on a sample basis and is the reason for why the required data needs to be pseudonymous.

F) Level of identification required
Pseudonymised - using an identifier that enables linkage back but where the data is for the researcher effectively anonymised - i.e. the key is held in the clinical domain or by an Honest Broker/ Trusted Third Party.

A level of geographic locality is required- region of country would be acceptable.

G) Dictionaries
Access to all dictionaries in current use within the clinical systems and mapping to older dictionaries that were used before.

Snomed CT (medical codes)
DM+D (Drugs, devices etc)

Mappings to:
- ICD 9/10 (medical codes- hospitals)
- READ codes (medical codes- GP)
- OXMIS codes (Pre READ system)
- MedRA (medical, adverse drug reaction codes)
- Existing drug code systems- BNF, ATC

H) Timings /data delivery/governance/pre-study
Active surveillance of drug use does not require 24/7 data delivery. Data delivered within a 7-10 day period would be ideal but even a 30 day window would be acceptable. However the volumes of data generated by such a system might suggest that a continual feed to pharmacovigilance server system might be ideal.

As pharmacovigilance is a public health issue it needs to be handled centrally for the principles of, how and what, but also cope with the specific study requirements of hypothesis testing studies.

Thus PIAG, Ethics and Caldicott approvals could be handled centrally and across the whole system whilst individual research studies would require their own scientific approval.

4.3 Advanced Prescribing decision support

A) Outline data needs
This requires feedback from data stored within the GP and hospital system as well as data from electronic transfer of prescriptions (ETP) and laboratory systems. It also requires linking to external data sources so that new warnings and Dear Doctor letters can be put in front of prescribers in the most timely and appropriate manner.

B) Sources of data

B.1	Primary	YES
B.2	Secondary	YES
B.3	Tertiary	YES
B.4	Pharmacy	YES dispensed data and pharmacist prescribing
B.5	Patient	YES, self reported ADRs.
B.6	Others Healthcare -state	Dentists, Pharmacists, Nurses, Private health centres who prescribe and those who administer POMs and those who administer medicines
B.7	Others – Social care	No
B8	Death	No
B9	Census	No
B10	Existing registries	Regulatory notices as well as the ability of a pharma company to voluntarily withdraw a product
B11	Others- state	

C) Temporal needs of the data
A real time system

D) Size/population of dataset required
All systems that enable prescribing.

E) Source data verification
Not applicable

F) Level of identification required
Not applicable- essentially a clinical system

G) Dictionaries
Not applicable other than within clinical system

H) Timings /data delivery/governance/pre-study
Not applicable

4.4 Other paradigms

The data requirements for other paradigms of medicine are to enable data streams from practitioners of each type to be incorporated into a fully integrated record or the facility for record linkage of these data. In some circumstances research could be conducted on the individual record itself (adverse effects related to use of acupuncture needles for example), in others an integration with the total record would be beneficial (the use of herbals and western medicines in relation to drug herbal interactions).

4.5 Disease and wider health surveillance

A) Outline data needs
The data requirements of the Health Protection Agency are considerable. They include:

- Patient identifiable data to enable the detection, tracking and management of infectious disease cases and their contacts.
- Notifications of clinical diagnoses and full laboratory reports from pathological diagnoses including 'suppressed' results such as complete antimicrobial panel results
- Symptomatic and syndromic surveillance to identify possible outbreaks of infectious disease or hazards to health
- Pharmacy data to enable analysis and interpretation of trends between prescribing behaviours and anti-microbial resistance
- Coverage statistics to determine the effectiveness of vaccination and screening programmes
- Linkage with hospital episode, death and other registry statistics to enable the detection of mortality and morbidity rates subdivided by a wide range of organisms, clinical disciplines and areas of interest
- Data to enable the detection and surveillance of hospital acquired and surgical site infections
- Surveillance and detection of risk factors and symptoms associated with chemical and radiation hazards
- Data from primary care, community and hospital services to inform and monitor screening programmes, eg genital Chlamydia screening

B) Sources of data

B.1	Primary	YES
B.2	Secondary	YES
B.3	Tertiary	YES
B.4	Pharmacy	YES, dispensed data and pharmacist prescribing
B.5	Patient	YES, self reported ADRs.
B.6	Others Healthcare - state	Dentists, Pharmacists, Nurses, Private health centres, Pathology
B7	Others Social care	No
B8	Death	YES
B9	Census	Probably Not
B10	Existing registries	Probably Not
B11	Others- state	NHS Direct

C) Temporal needs of the data
The Health Protection Agency has a number of surveillance systems currently in place. However for SUS to provide an effective service retrospective data will need to be included or data will have to be downloadable to HPA based systems. Some data will be required in real-time timescales.

D) Size/population of dataset required
The entire population will need to be monitored for the majority of the surveillance functions of the HPA. Some areas of surveillance only require information on specific sub-populations however e.g. age ranges or diagnoses of a specific infection.

E) Source data verification
It is assumed that the SUS only contains verified and accurate data. This is essential.

F) Level of identification required
Much data required needs to be patient identifiable. Subsets of data could be pseudonymised, identifiable only by soundex, age range, gender and/or postcode. Some subsets can be aggregated data.

G) Dictionaries
The HPA uses a range of proprietary coding systems. It is expected that use of SNOMED CT and DM+D will be sufficient.

H) Timings /data delivery/governance/pre-study
Much data requires near real time delivery. The HPA has section 60 exemption of the data protection act enabling the use of the patient identifiable data for specific named purposes.

5. Key issues emerging

Key to VISION delivery are:

Technical

1. Band width to enable the delivery of **all** patients' **full** electronic record, code and text from all parts of the care delivery system to a central system.
2. Computational capacity to meet analysis requirements.
3. GRID or large scale enterprise technologies are emerging that will address limitations through bandwidth and computational power.
4. A storage system for such data that enables rapid access to the relevant data in order to assemble the required data for cohorts of patients taking certain medications or who have certain disease or symptom characteristics and to assemble the controls who match those patients.
5. Enablement of the required decision support systems across all clinical systems and all parts of the NHS including contracted services.
6. Use of clinical coding and dictionaries in a universal and standardised manner across all parts of the NHS.
7. Mapping of clinical coding systems to specific coding systems used in various aspects of pharmacovigilance – e.g. MEDRA , BNF and ATC.
8. Use of tools and methodologies appropriate to the high end requirements of data access, data manipulation, data modelling, the needs of control selection and of case control, cohort and other pharmacoepidemiology studies. There are a few currently available tools for either the data manipulation or the statistical end of the work.
9. The availability and employment of people with suitable large dataset skills together with the required expertise in pharmacoepidemiology.

Data issues

1. The data currently in SUS do not enable pharmacovigilance to be fully or even in part achieved.
2. Access is required to as much historic data as can be made available. Historic data is required to undertake, to acceptable levels, most hypotheses testing research.
3. Access to the full personal level GP record is required - code and text in a suitably anonymised form that enables validation - including data in sealed envelopes. The drugs used for conditions that might be included in the sealed envelope are not exempt from drug safety issues.
4. Access to hospital bed, clinic and day case prescribing and other treatments as well as diagnosis/symptoms in the same way as with GP records, laboratory data and other specialist test data.
5. Access to the dispensed personal record via ETP and prescribed and dispensed record of pharmacists where they are acting as an independent prescriber.
6. Access to data from care given by 'carers' contracted by the NHS.
7. Access to prescribing and other relevant data from all independent prescribers.

Temporal issues

Real time access to electronic record data is required for disease surveillance and incident reporting using specified systems for all types of events. Pharmacovigilance requires data feeds from electronic records within days/ a few weeks.

Expert Pharmacovigilance decision support systems, as recommended, would need to be incorporated at the point of clinical care delivery and therefore have access to the patient record at that time.

Governance

Scientific governance and review of all research to be undertaken is an absolute requirement to ensure important public health safety work is never compromised.

The ownership, funding and overall responsibility needs to be addressed.

Communication with patients and healthcare professionals and other stakeholders must be handled

sensitively.

The use of existing models of consent and anonymisation that have been proven by the test of time (GPRD and MEMO/Tayside) - implicit consent with patient opt-out for the use of, essentially anonymised data by the researcher, but with a way back for validation where the key is held in the clinical domain or by a trusted third party/ honest broker.

6. Summary

NHS CRS in its entirety has the ability to support an advanced pharmacovigilance system that, like the current UK system, is a world leader. Pivotal to this are the benefits delivered by the NHS cradle to grave cover with one electronic healthcare record. This and the advanced stage of delivery provide added advantages over other e-healthcare systems that are in development (USA, France, Germany) or indeed already there (Denmark and Sweden).

- The datasets currently available within the Secondary Use Service (SUS) do not meet the necessary requirements.
- Text as well as coded data is required.
- The data from full clinical records is required including that proposed in sealed envelopes.
- It should build on what is already available and is accepted as gold standard in pharmacovigilance.
- Data at the research level should be anonymised but there needs to be a way back for the purpose of validation. Such a system can be enabled by SUS acting as the 'honest broker' and holding the pseudonymisation key.
- Decision support systems capable of immediate update across all parts of the web enabled NHS need to be provided.
- Patients should be encouraged to contribute data related to relevant events via their Health space and an automated onward system to the correct authority.
- Data from practitioners of other types of therapy should be incorporated within a patient's record if at all possible.
- An NHS wide standard incident reporting system, for other than adverse drug reactions that have an accepted path to the MHRA, with adequate coding to distinguish types of incidents so they flow to the correct authority is required.
- GP systems should enable suspicion reporting related to devices.
- Support to improving efficiency in delivery and implementation of Risk Management Planning for the pharmaceutical industry.
- Disease surveillance requires real time reporting and in some cases reporting of 'suppressed data'.

Appendix 1 MHRA Yellow Card

COMMISSION ON HUMAN MEDICINES

In Confidence

SUSPECTED ADVERSE DRUG REACTIONS

MHRA

If you suspect that an adverse reaction may be related to a drug, or a combination of drugs, you should complete this Yellow Card or complete a report on the website at www.yellowcard.gov.uk. For *intensively monitored medicines* (identified by ▼) report **all** suspected reactions (including any considered not to be serious). For *established drugs* and *herbal remedies* report **all serious** adverse reactions in adults; report **all serious and minor** adverse reactions in **children** (under 18 years). You do not have to be certain about causality: if in doubt, please report. Do not be put off reporting just because some details are not known. See BNF (page 10) or the MHRA website (www.yellowcard.gov.uk) for additional advice.

PATIENT DETAILS Patient Initials: _____ Sex: M / F Weight if known (kg): _____

Age (at time of reaction): _____ Identification (Your Practice / Hospital Ref.)*: _____

SUSPECTED DRUG(S)

Give brand name of drug and batch number if known	Route	Dosage	Date started	Date stopped	Prescribed for

SUSPECTED REACTION(S)

Please describe the reaction(s) and any treatment given:

Outcome

Recovered ☐
Recovering ☐
Continuing ☐
Other ☐

Date reaction(s) started: _____ Date reaction(s) stopped: _____

Do you consider the reaction to be serious? Yes / No

If *yes*, please indicate why the reaction is considered to be serious (please tick all that apply):

Patient died due to reaction ☐
Life threatening ☐
Congenital abnormality ☐
Involved or prolonged inpatient hospitalisation ☐
Involved persistent or significant disability or incapacity ☐
Medically significant; please give details:

* This is to enable you to identify the patient in any future correspondence concerning this report

Please attach additional pages if necessary

Please list other drugs taken in the last 3 months prior to the reaction (including self-medication & herbal remedies)

Was the patient on any other medication? Yes / No If *yes*, please give the following information if known:

Drug (Brand, if known)	Route	Dosage	Date started	Date stopped	Prescribed for

Additional relevant information e.g. medical history, test results, known allergies, rechallenge (if performed), suspected drug interactions. For congenital abnormalities please state all other drugs taken during pregnancy and the date of the last menstrual period.

REPORTER DETAILS

Name and Professional Address: _____

Post code: _____ Tel No: _____

Speciality: _____

Signature: Date:

CLINICIAN (if not the reporter)

Name and Professional Address: _____

Post code: _____

Tel No: _____ Speciality: _____

If you would like information about other adverse reactions associated with the suspected drug, please tick this box ☐

If you report from an area served by a Yellow Card Centre, MHRA may ask the Centre to communicate with you, on its behalf, about your report. If you want only MHRA to contact you, please tick this box. ☐

Send to **Medicines and Healthcare Products Regulatory Agency, CHM FREEPOST, LONDON SW8 5BR**

Yellowcard

Annex A Section 2

UKCRC Connecting for Health Research Simulation Report: Clinical Trials

Mr Rob Thwaites, GlaxoSmithKline R&D
Professor Nish Chaturvedi, Imperial College
Dr John Logie, GlaxoSmithKline R&D
Dr Lisa Davenport, Johnson & Johnson
Dr John Parkinson, MHRA (GPRD)

1. Introduction
2. Benefits of clinical trials research
3. Current situation and vision
   a) Clinical trials research in the UK now
   b) Clinical trials research in the UK in the future
   c) Simulation findings – needs highlighted by the protocol assessment
   d) Simulation findings – needs identified from description of clinical trials research in the future
4. Specific data needs
5. Key issues emerging
   a) Major issues
   b) Implications for the way forward
6. Summary
7. Appendices

   Appendix 1: Report from the protocol assessment of the ADOPT study (summary)
   Appendix 2: Table listing the specific information about the data elements required for a 'minimum data set'
   Appendix 3: Summary of the standard types of data collected during a clinical trial, organised in the way required by the FDA
   Appendix 4: Comparison of clinical trial data requirements for regulatory agencies with data types available in GPRD

1. Introduction

The purpose of the simulation is to explore the capability of the NHS Care Records Service (NHS CRS) infrastructure to support clinical trials research. As the NHS CRS is not currently populated with the data needed for clinical trials research, this simulation is based in part upon a protocol assessment of a clinical trial, using data currently available rather than data that would be available once the NHS CRS is complete. It is also based in part upon a review of how clinical trials are conducted, in order to explore the implications for future data needs and associated requirements for the NHS CRS.

Results from protocol assessments provide an early warning of potential recruitment issues that may arise in searching for subjects meeting the trial inclusion and exclusion criteria. The clinical trial that has been assessed here is the ADOPT study[1], conducted in primary care, with the assessment based upon data from the General Practice Research Database (GPRD)[2]. The results of this ADOPT protocol assessment have been used in this report to highlight technical, operational and governance issues that will need to be addressed if the NHS CRS is to be designed to support protocol assessments in the future.

The subsequent description of the conduct of clinical trials is extended to explore the key implications for NHS CRS for conduct of interventional clinical trials in the future. For the purposes of this exercise, the specific type of clinical trial used in the protocol assessment and described in the simulation below is a pharmaceutical company sponsored, randomised, clinical trial conducted in a primary care setting. The relevance of the issues drawn out to other types of clinical trials is discussed in Section 5.

There is already a wealth of relevant experience within the UK of developing specific or local systems and capabilities to support clinical research. In the time available to conduct the simulations, it was not possible to review widely or systematically the work

1) ADOPT (A Diabetes Outcomes Progression Trial) is an international, multi-centre, randomised, double-blind study involving 4,360 drug-naïve people who had recently been diagnosed with type 2 diabetes at sites in the UK and elsewhere.
2) The GPRD is the world's largest database of anonymised longitudinal medical records from primary care, owned by the UK Medicines and Healthcare Products Regulatory Agency.

that is already underway around the country. The simulation team was able to explore the possibilities for the future through discussions with medical and informatics staff in Manchester and in Scotland where good progress is already being made, and examples from work there are referred to in the report.

2. Benefits of clinical trials research

The UK has a significant opportunity to increase clinical trial activity through the successful development of the NHS CRS. Clinical trials bring important benefits to the UK: to patients, to academia, and to the economy. For pharmaceutical and other healthcare companies, choice of location for clinical trials is based on a number of factors, including for example, availability of investigators and patients, the regulatory environment and costs. While the UK has traditionally been a leading location for clinical trials research, the level of activity here in recent years has suffered as newer, lower cost sources are now taking an increasing share of global clinical trial activity. With the NHS CRS is in place, however, clinical trials could be conducted more efficiently, more predictably and more cost-effectively, and the UK will be a much more attractive place for locating clinical trials research.

Clinical research in the UK is substantial and brings benefits to patients in general through better detection of illness, improved treatments and improved outcomes.. Clinical trials activity is a major component of clinical research in the UK, bringing potential benefits to patients through encouraging earlier and broader use of new technologies. This includes medicines, devices and diagnostics. More specifically, participation in clinical trials can also bring benefits to those patients taking part through closer medical attention and continuity of care during and even after the trial.

Academic and other research institutions benefit from the location of clinical trials in the UK through funds and investment into research centres. Through UK clinical trials, or global clinical trials that include sites in the UK, individual researchers have the opportunity to work with the latest developments in treatment and this in turn enriches the quality of scientific research in the country. In the longer term, with richer research opportunities and continued funds, the UK is better able to retain researchers and their skills and knowledge in a globally competitive market.

Location of clinical trial activity in the UK also brings enormous economic benefits. The report[3] produced by McKinsey & Company in 2005 estimated that of a total annual expenditure of £6 billion in 2003, industry spent £4.2 billion on clinical research with the pharmaceutical industry contributing about £3.3 billion of this. While no breakdown of this by type of clinical research was provided, given the typical proportion of pharmaceutical companies' research expenditure on full development activities, the extent of clinical trial activity might amount to £2 billion to-£3 billion per annum.

The UK is still a leading centre in Europe for clinical trial activity (more commercial clinical trials were conducted in the UK than in any other major European country in 2005[4]) and has a significant pool of talent and infrastructure for growth. However, while there has been some, albeit small, increase in clinical trial activity in the UK over recent years, this has been overshadowed by the significant increases in activity in newer, lower cost countries, notably in Eastern Europe and Asia.

Offering lower costs and newer sources of patients, countries such as Russia, China and India have attracted an increasing share of global clinical trial expenditure over the last 5 years, notably for the large, multi-million pound, later stage clinical trials. To compete in the future for a larger share, the McKinsey report asserted that the UK will need to be competitive in five areas: strategic relevance, quality, time, reliability and cost; but warned that

3) Commissioned by the UKCRC, McKinsey & Company produced the report Clinical research in the UK: towards a single system that reliably delivers distinctive quality and rapid access at reasonable cost in August 2005
4) see report by McKinsey & Company, Chapter 3
5) see report by McKinsey & Company, page 19

in the last three of these the UK was below average internationally, and only 'on a par' for the first two[5].

Encouragingly, however, the McKinsey report concluded that a key component in creating competitive advantage for the UK in the last four of the five areas above would be the comprehensive and flexible healthcare IT system from NHS CfH noting that 'Connecting for Health could…create the world's largest integrated and shareable patient record IT system'[6]. More recently Cooksey highlighted the unique opportunity offered by the NHS CRS, commenting that: 'First and foremost is the potential offered by the new Connecting for Health IT database…'[7] in making the UK attractive as a location for clinical trial activity by industry. There is a clear belief that with a flexible NHS CRS in place, clinical trials could be conducted more efficiently, more predictably and more cost-effectively, and the UK will be a much more attractive place for locating clinical trials research in the future.

3. Current situation and vision

3a. Clinical trials research in the UK now

A wide range of clinical trials are conducted in the United Kingdom, many of them forming part of larger, multi-country trials. While this range encompasses a variety of designs, interventions and measures, the underlying process is similar for each: detailed planning as the cost of a trial can run into millions of pounds; recruitment of sites and subjects; conduct of trial including data collection; analysis and reporting of results; and, on occasions, post-trial work.

At the time of *clinical trial planning*, a number of studies are done to finalise the design and plan for a study. These include:

- profiling a patient population prior to initiating a clinical trial, to better understand the natural history of the disease within the population under study
- informing study design, e.g. a study of the expected rate of an outcome to inform statistical power calculations and therefore the number of subjects required
- assessing the feasibility of recruitment

For a pharmaceutical company-sponsored study, for example, most of this work would be done by the company with expert guidance from a limited number of external (ie non-company) clinicians, some of whom may eventually participate in the study as investigators.

For a third of these activities, the company could undertake a general assessment of the protocol, where the proposed eligibility (inclusion and exclusion) criteria defining the subjects to be entered into the clinical trial are applied to real-world data on patient populations to assess the numbers of subjects available, and to explore the likely impact of adjusting individual criteria on expected recruitment. Where this is done, it would be explored using an existing (typically limited) electronic patient database, but this would be the minority of cases; far more common would be a survey of potential recruitment sites.

With the latter approach, potential sites would be involved, with the company sending the protocol and asking the site to undertake a local feasibility study, including responding with an indication of how many subjects meeting the eligibility criteria the site could find in the recruitment time frame. At the site, the staff pull all case notes out on individuals that might be suitable, or they may look through patient details on a stand alone register if one exists, and would then estimate how many patients they could recruit.

With this manual approach, estimates are often inaccurate, and typically over-optimistic. In addition, this approach is time-consuming and often the time lines given to the sites to respond with potential subject numbers are so short that they do not even permit a manual trawl of the patient notes. Incomplete chart data and the inability to link to supporting information about patient history or co-morbidities, to categorise groups of patients or to generate risk profiles, leads to educated guesses, at best, which is a costly exercise for all involved.

Within the pharmaceutical company, the protocol

6) see report by McKinsey & Company, page 30
7) in the Foreword by Sir David Cooksey to A review of UK health research funding, December 2006

would then be finalised, as would additional documents, including (for example) case report forms (CRFs) used for data collection; the Clinical Trial Agreements and Study Guidelines outlining the roles and responsibilities of the staff at the sites recruiting subjects for the trial; and associated patient documents such as the informed consent form and letters of invitation for participation.

Within clinical trial planning, activities associated with clinical trial design can be carried out using electronic patient level data, and much of the evaluation can be done using this data in an anonymised form. Currently, however, where assessment of potential subjects is typically conducted by hand at a site, access to personally identifiable information is now needed. Where this assessment can be done electronically, the need for personally identifiable information is eliminated.

For the *recruitment* phase of the trial, the identification of investigators can be targeted by identifying 'pockets' of the disease from epidemiology data. Existing databases built on electronic patient records could assist with this, but in practice are rarely used for this because of their incomplete coverage.

Once the relevant approvals for the study are received and the sites have been initiated and trained, identification of patients who could become trial subjects then begins. This can be based on advertising or contacting patients already known to the investigator, but one route for this based on patient records is for the investigator to arrange for a nurse or study co-ordinator to go through the site's patient records identifying individuals who might be eligible. Given that primary care in England is largely computerised, some of this work can be done with the electronic records for primary-care based trials. The coding required to tackle this in detail, however, is significant and in practice this can only be done in a limited way. As a result a significant proportion of potential subjects still fails in the subsequent, detailed screening process. In addition, in secondary care, patient records remain largely paper based and screening for secondary-care based studies is inevitably labour-intensive.

Once potentially eligible subjects have been identified, they will be contacted for a pre-screening assessment, usually, but not always, at an appointment. If pre-screening is successful, the patient will be invited to attend clinic for a screening visit. At this time, the written informed consent sought would also include agreement from the subject for persons connected with the study to be able to access the subject's (personally identifiable) health care information. With consent in place, and the subject's screening visit successful, their GP is notified of their participation in the trial.

In studies with detailed eligibility criteria, screening is a very time-consuming and labour-intensive exercise. The investigator or related staff (and subsequently the study monitor responsible for checking that screening has been done correctly) will need to go through pages of historical notes, organised chronologically, in order to verify that the subject meets all the criteria.

Within the clinical trial phase, there are numerous activities that require someone on the research team to be able to access patient-identifiable information.

During the data collection phase as the trial is being conducted, there has been a significant shift in recent years in the amount of trial data captured electronically (known as electronic data capture, or eDC). For pharmaceutical company sponsored clinical trials for example, as much as 40% of trial data is now captured electronically. There is, however, little emergence yet of the transfer of existing data, entered into or already residing in an electronic healthcare records, directly into the clinical trial eDC system.

Source data verification is a critical and required[8] component of ensuring the quality of data capture; the process for source verification is changing as more data are captured electronically. For the information that is still recorded on a paper CRF, the study monitor (who could be a pharmaceutical company employee) verifies a statistically controlled

8) Source Data Verification is a requirement from ICH E6: Good Clinical Practice

Table 1 Summary of clinical trial activities (without post-trial follow-up)

Typical process for a pharmaceutical company sponsored clinical trial



Study design and planning	Initiation and recruitment	Study conduct	Analysis and reporting
Need for specific CT identified in the development plan Concept protocol written Possible sites surveyed for estimates of eligible subjects Analysis plan agreed Clinical trial protocol, plan and budget finalised Trial registered	Sites/investigators identified Approvals received from appropriate committees Sites initiated and trained Subject recruitment programme undertaken Potential subjects identified; consent requested GP notified of subject participation Subjects screened Subjects entered into trial and allocated code	Subjects follow requirements of protocol Study staff collect data using CRF Subjects' safety monitored Sponsor conducts monitoring visits Sponsor conducts source data verification Sponsor receives (coded) data in-house	Sponsor 'cleans' then locks databases Sponsor analyses data Sponsor writes final study report Sponsor and investigators disseminate results

sample of data recorded with source data in the patient's record, depending on the objectives, complexity and size of the trial.

During the data collection phase, there are many activities that require access by a member of the research team on site to patient-identifiable information. Patient consent will have been given for this at the start of the study.

Data will be needed for *post-trial* work that may be conducted following completion of the trial:

- Extension studies – the ability to follow trial subjects after the end of the study, to assess long-term outcomes
- Additional studies, for example to meet post-marketing commitments, including studies to assess the real-life outcomes of a treatment following approval (such as efficacy and safety outcome studies and pregnancy registries).

While there are examples of electronic health records assisting in several aspects of clinical trials research outlined above, the use of such approaches currently is limited, sporadic and piecemeal.

3b. Clinical trials research in the UK in the future

For *clinical trial planning*, national electronic patient records can be used to investigate issues during the design of the trial and to increase the success of recruitment through accurate feasibility assessment by:

- informing study design, e.g. a study to understand prevalence of co-morbidities in the population to be studied, or a study of the expected rate of an outcome to inform statistical power calculations and therefore the number of subjects required
- more accurately and quickly assessing the likely feasibility of recruitment given the eligibility criteria by conducting protocol assessments using the electronic, patient-level data
- identifying investigators or sites where recruitment of subjects can be managed and data collected from subjects

In general, the clear advantage of a national patient record is that these studies or investigations will be more thorough, more accurate and therefore more reliable than existing approaches. Many or all of the

elements of the protocol eligibility criteria can be selected for every person on the national register. Not only does this increase the number of patients that can be assessed but it will increase accuracy of subsequent selection and reduce the number of screening failures.

Without the need to trawl through chronologically-organised and sometimes very lengthy hand written case notes per subject, identification of suitable patients as potential subjects becomes more accurate. With a search and select electronic facility, the user selects only the information that is relevant and has no need to review the rest of the unrelated record. Inevitably, this gives a more accurate, and far quicker, assessment. There are already examples of where this has been undertaken, though on a smaller scale, in Manchester and Scotland, or using large primary care databases (as in the protocol assessment in this report).

Rather than searches on a local level in local populations, a national patient record will enable one central search, thereby vastly reducing the number of man hours required to complete the process in several geographical locations.

As the feasibility process is not just to identify suitable patients, but to confirm availability, suitability and interest from investigators and staff, once geographical locations with sufficient subject numbers are identified, the corresponding Research Networks or NHS Trusts will be approached to confirm interest in the study and provide suggestions for suitable local research teams.

Furthermore, the electronic health record would allow for data-driven decision-making before the trial begins, about the extent and timing of source data verification. Algorithms based on the design, complexity and size of the trial could be applied to the eHR and the results used to establish statistic and logistic parameters for audit. This would assist not only in safety, compliance and resource planning, but also provide feedback to the medical and informatics communities about gaps or inconsistencies in eHR data collection or clinical trial design.

Time is of the essence when developing a new medicinal product, so when using a national electronic record, not only is the resource required greatly reduced and accuracy increased, but the speed with which accurate feasibility assessments can be delivered is significantly reduced.

For the *recruitment* phase of the trial, electronic patient records can be used to increase the success of recruitment by:

- assisting the investigator in identifying potential subjects to be invited to participate in the trial and reducing the numbers of screen fail subjects
- allowing individuals themselves to flag their interest in participating in clinical trials by recording in their Health Space record their willingness to be approached to take part
- generating a saving of time and costs with the shift in emphasis of subject recruitment away from advertisements
- greatly reducing resource required for the initial screening, thereby permitting more time for study related staff to dedicate to actual trial subjects and visits, with, ultimately, an increase in subjects recruited per centre.

Use of a detailed and reliable electronic health record will reduce significantly the number of screening failures. Failures are expensive, not only in terms of the direct costs (expenditure) and indirect costs (opportunity costs associated with delays in completing the study) of the clinical trial, but also the societal costs, including lost time or working days for the medical personnel and for the potential subjects.

Even for those subjects who are ultimately entered into the trial, the costs associated with screening will be much lower with an electronic health record in place. Rather than manual screening by the investigator, and subsequently the monitor, use of the electronic record would permit rapid searches and automated assessment, in order to verify that the subject meets all the criteria. Assuming a complete electronic health record is in place (including standard controlled terminologies), electronic chart review would reduce risks to patients with potential

contraindications for study medication or procedures that may have been overlooked in manual review, by automatically excluding them from the subject pool.

At a research site, this in turn will permit the research staff to devote more time to increasing patient numbers entered into the trial and the care of ongoing subjects. A move to increasing subject numbers per research centre is advantageous for the UK as this is an area in which the UK is less competitive compared to other countries that participate in clinical trials.

National searches will widen the catchment area for research sites, with patients from outlying areas being included in the search for subjects. Currently, this only happens on a referral request basis at a specific site, with requests being sent manually to surrounding GPs or other consultants in a secondary care setting. This is time-consuming and relies on voluntary assistance from professionals outside the research network. If subjects within an acceptable travelling distance could be identified automatically using the national record, not only will more people have the opportunity to participate in clinical research, but the number of subjects per centre is increased.

Use of a national patient record would negate the need for advertisements for clinical trials. In whatever format advertisements take, they often include only basic information regarding the disease area and top line eligibility criteria. Therefore companies or other trial sponsors need to employ additional personnel or contract call centres to pre-screen these patients with a more in-depth questionnaire covering the more detailed entry criteria. This is a time-consuming and expensive way to identify possible subjects. By having access to the national electronic record, the sponsor would be confident that the process was identifying all suitable subjects using a strict set of inclusion and exclusion criteria from the outset.

The recruitment phase is typically the first time identification of patients will be necessary. Prior to this only anonymised electronic patient-level data would be needed. Although the advantages of a national electronic database for the purposes of clinical trials recruitment (and later phases) are clear, there will need to be systems and safeguards to ensure that the rights of all individuals continue to be respected.

During the *data collection* phase as the trial is being conducted, electronic data and systems can increase the efficiency and accuracy of data through, for example:

- Electronic capture of trial data - clinical trial data can be entered directly into the electronic health record and transferred to the sponsor as required
- Remote and secure source data verification by the trial sponsor or monitor
- Offering safety checks for ongoing subjects in the trial

Using the electronic health record as the point of source data has several advantages over the current paper based system. Even if were not possible for the electronic data from the national record to be transferred automatically into the electronic CRF, the needs for on-site source data verification (a cross check between what is recorded in the health record and what has been transcribed to the electronic CRF) is reduced as monitors are able to access both systems remotely.

A significant proportion of the total cost of running a clinical trial can be taken up by on-site monitoring, as companies or other trial sponsors send out their monitors to assess research centres country-wide. As the need for this on-site work is reduced and eliminated with the use of electronic health records, there are significant benefits including reduced travel and staff financial costs for the sponsor and less resource required by site staff in assisting in data verification.

Should a comprehensive IT solution be available for automated direct transfer of electronic patient data to the electronic CRF, then provided the systems were validated and backed-up, source data verification may not be necessary at all. Even informed consent could be performed by electronic signature, also improving patient privacy with the name and personal details of the subject not being a requirement for verification. For direct transfer to

work, all patient data would be required to be on the electronic record in a transferable format, including but not limited to ECG scans, laboratory data, X-rays, and prescribing data.

Furthermore, in terms of patient privacy, with paper based records, the patient provides consent to their full details being accessed and reviewed by personnel involved with the study, both from within the Health Service and by external groups such as the pharmaceutical monitor and regulatory authority inspectors. Using an electronic record would mean that the full name and contact details of the patient would not need to be revealed and the data verification could be performed using only the patient identifier.

Often for a clinical trial, a subject is requested to record information about their disease or other trial related requirements on a regular basis using a diary or PDA. It will be possible for the patient to enter information directly, either by accessing the appropriate part of the electronic record or by automated transfer from a suitable PDA device for example.

Indeed, with suitable IT solutions, the national record could even flag up issues that the current method of trial data collection would fail to see, thereby further improving patient safety. Two examples are:

- If a patient were admitted in an emergency situation, the record would alert the medical professional user to the fact that the subject was in a clinical trial and would provide a mechanism by which the research code could be broken if medical assessment dictated that was in the best interests of the patient
- The majority of clinical trials include a list of forbidden concomitant medications due to possible drug interactions, leading to potentially harmful side-effects. With the inclusion of electronic prescribing data, the system could flag to the user that medications were forbidden for this subject as part of his/her participation in a clinical study.

Data will be needed for post-trial work that may be conducted following completion of the trial

- Extension studies – the ability to follow trial subjects after the end of the study, to assess long-term outcomes. Recent experience, for example, in Scotland with the use of electronic health records to follow up patients from the original WOSCOPS study, demonstrated how much more effective and straightforward this could be than is currently the case[9];
- Additional studies, to meet post-marketing commitments, including studies to assess the real-life outcomes of a treatment following approval (including, efficacy and safety outcome studies, pregnancy registries etc)
- Safety studies initiated in the event that long term side effects appear to be associated with use of a product

Irrespective of the need for contacting subjects post trial, the process is clearly more efficient when a real time electronic health record is maintained for each subject nationally. Currently, if a patient were to relocate after completion of the clinical trial, the process of contacting the patient often fails due to inadequate links between paper based records.

The records will need to be robust to system and other changes that occur over time. For paediatric research, there is a need to follow children up over a very long period of time, and the information recorded would need to be robust to various changes, for example in NHS organisation or data management systems.

3c. Simulation findings – needs highlighted by the protocol assessment

The protocol that has been assessed here is for the ADOPT clinical trial, conducted in primary care, with the assessment based upon data from GPRD. While the usual purpose of a protocol assessment is to

9) Ford et al were able to use electronic health records, linking to secondary care records and mortality records, to follow up participants that had enrolled 15 years before in the West of Scotland Coronary Prevention Study (WOSCOPS) (a recently-completed study quoted in Focus on Research, Follow-up of WOSCOPS subjects: long-term benefits of treatment with pravastatin and relationships of baseline inflammatory markers with fatal and non-fatal cancers issued by the Scottish Executive Health Department Chief Scientist Office)

generate results which will provide an early warning of potential recruitment issues that may arise in searching for subjects meeting the trial inclusion and exclusion criteria, in this case the objective was to highlight broader issues. From the conduct of the assessment, it has been possible to identify several technical, operational and governance needs that will have to be addressed if the NHS CRS is to be designed to support protocol assessments in the future.

The primary objective of the ADOPT trial was to compare the effects of long-term treatment on the improvement of glycemic control in patients with recently diagnosed type 2 diabetes mellitus. The primary inclusion criterion for the trial required patients to have had diabetes for ≤ 3 years, to be drug naive, and with a pre-screening fasting plasma glucose (FPG) level of 126-240 mg/Dl. General practice records, from the UK GPRD, were used to assess the impact of these and other eligibility criteria.

In the assessment, patient histories were searched to identify records of diabetes. With longitudinal records, it was possible to establish if these reports represented new diagnoses, or the continuation of existing disease. With the inclusion of prescribing histories, it was possible to establish if patients had received prior treatment for glycemic control. Both illustrate the importance of historical and linked data.

Current (managed) disease may be captured from administrative systems containing medical coding. Within this assessment, however, patients' medical histories were required to inform key trial criteria. For example, patients were deemed ineligible if they had a history of angina, congestive heart failure, anaemia, lactic acidosis, renal or hepatic disease etc. Whilst some of these conditions could potentially be assessed based on current data using explicit medical records, or surrogates, discrete historical events would not be captured. In this study, for example, hysterectomy status was reviewed to identify patients of childbearing potential, demonstrating the need for historical, as well as ongoing longitudinal, data.

The importance of early identification of patients from GP records is illustrated by the relative importance of the key inclusion criteria. Of all type 2 diabetes patients identified, 71% had been diagnosed more than 3 years previously; more than 96% had received prior drug treatment. Recently diagnosed diabetes patients, without drug treatment, are likely to be managed within general practice, and potential trial subjects may not have been identified by their use of other services (such as prescription or referral services etc). Only access to primary care electronic data would allow these recently diagnosed diabetes patients to be included in the protocol assessment, and subsequently identified as possible subjects for recruitment.

Additionally, many of these conditions may also have been treated or diagnosed within secondary care. Secondary care information is limited in primary care datasets, such as GPRD, to the data the GP deems clinical relevant to record for a patient's on-going management, and therefore chooses to enter into their GP records. The completeness of information would be improved by linked data recording from secondary care settings.

Other key criteria required test results or risk factor and lifestyle information. Fasting Plasma Glucose level was a primary inclusion criterion for this study while both elevated blood pressure and current alcohol/drug abuse were exclusion criteria. Laboratory test information, a prerequisite for many protocol assessments, may be captured elsewhere in the health system. Other patient assessments, such as blood pressure or BMI, however, will only be captured in GP records. The requirement for these data demonstrates the need for an integrated system, including laboratory and GP records, and linked by a unique patient identifier.

A key point within the protocol assessment is that all eligibility criteria are applied at the individual patient level. These analyses could not be conducted using population estimates of disease prevalence, since these would not account for the co-occurrence of multiple co-morbidities within an individual patient. More generally, the need to consider multiple combinations of factors (reflecting the multiple criteria that need to be satisfied for eligibility), requires access to patient-level data for the analysis.

The assessments also need to be based on 'clinical' information, rather than surrogates, such as administrative records. For example, if 'medical' information is to be based on use of services (e.g. referrals to specialist care), it has to be clear that a referral truly indicates the presence of disease, rather than a consultation where a patient's disease status will be assessed. This is essential to avoid over-estimates of disease prevalence that would inflate the estimates of the total number of potential subjects, and distort the impact of individual eligibility criteria. A record designed for one purpose may prove to be misleading for a secondary purpose.

A limitation of the data currently available from primary care databases such as GPRD is that the figures are based on recent (perhaps 6 months old), rather than current information. This could be improved under the current arrangements and would be a desirable feature for protocol assessments and a necessary feature for clinical trial recruitment.

Despite the fact that the availability of primary care databases in the UK is world-leading, only a proportion of the population (typically less than 5%) is included in each. The strictness of eligibility criteria in a clinical trial can reduce the final set of subjects shown to be available through the protocol assessment to a very small number. The subsequent work that is done, in fine-tuning the eligibility criteria and then re-running the protocol assessment would be greatly enhanced with a database, say, ten times the size of GPRD.

3d. Simulation findings – needs identified from description of clinical trials research in the future

Within the range of activities involved in running a successful clinical trial in the future, there will be a need to access and process data from electronic records at a number of stages before, during and after the trial. At different stages, a range of data elements will be needed, though these elements need not always be personally identifiable - anonymised or pseudonymised would be sufficient, depending on the purpose. These include:

- Access to the subject's complete primary care record, eg for protocol assessment (anonymised), for screening by the investigator (personally identifiable), during data collection (pseudonymised). The level of detail provided in the currently available primary care datasets, such as GPRD, is acceptable.
- Historical, subject-level information from the healthcare record(s), in some cases up to 10 years' data, eg for protocol assessment (anonymised) and for screening by the investigator (personally identifiable data). The extent of historical data available through the currently available primary care datasets, such as GPRD, is acceptable.
- Access to detailed information about medication use, including information on drug, dose, and method of administration – the information contained in the ETP system should be adequate for this. This information will be needed, for example, during screening by the investigator (personally identifiable data) and for recording concomitant medication use (pseudonymised data). For many purposes, it will be important to be able to link the medication with a diagnosis. The information to be contained in the ETP system should be adequate in terms of detail, though historical data will need to be available and linkages across databases effective.
- Access to detailed laboratory test data, including the test result and normal ranges, for screening by the investigator (personally identifiable)
- Access to the subject's complete hospital care record, This should include not only information currently available through the Hospital Episode Statistics database, but also detailed information about medication use and test information from the hospital setting.

In addition, as data will be housed in different systems, in order for these to be pulled together a unique patient code will need to be used. Systematic use of the NHS number could ensure a unique code for each patient. In combination with a central linkage facility, provided by SUS for example, this could provide the means for linking and, where necessary, aggregating data at the patient level.

Learning from the examples in Scotland and Manchester the simulation team was able to see the possibility of engaging local support for collecting and using the data.

In the Manchester area, small (quarter of a million patients) but comprehensive and integrated databases have been successfully established with the support of the local PCTs and Trusts, notably because these provide information of value to medical professionals locally. The local health intelligence hubs (a combination of the integrated database and the capability to create and analyse the data) are emerging as a result of benefits not only to research, but also to public health and, perhaps most significantly, to the local healthcare community.

In Scotland, a similar approach has led to a comprehensive and integrated set of databases in the Tayside area, and activities to link this to Scotland-wide data sources are now under way. Future work in the development of the NHS CRS should review the progress made in Scotland, not only for the lessons that have already been learned there that can be applied in England, but also for how future developments in England can be compatible with progress in Scotland and Wales.

4. Specific data needs

Summarised in the first part of this section is the list of data needs in the format request, and adopted in the 'Surveillance' simulation (sections A-H below). More detail on the specific data elements that would need to be captured for clinical trials research is included in the Appendix. Comments on the usefulness of the proposed SUS data are then given at the end of this section, based on an understanding of the data elements that are currently proposed to be available via SUS.

A) Outline data needs	
A.1 Real world-Clinical	Real world-Clinical
A.1.1 as recorded in existing systems	Yes, for trial planning, recruitment and data collection
A.1.2 Interventional could be recorded in existing systems	Yes for electronic data capture of clinical trial data
A.2 Special screen-clinical	
A.2.1 in existing system recording	Yes, for electronic data capture of clinical trial data
A.2.2 external portal recording	No
A.3 Patient/ non clinical recording	
A.3.1 Into existing system	Yes, would be useful for electronic data capture of patient-reported information
A.3.2 Into new system	
A.4 Other- state	

B) Sources of data		
B.1	Primary	Yes
B.2	Secondary	Yes
B.3	Tertiary	Yes
B.4	Pharmacy state POM, P, GSL	Yes (ie dispensed data)
B.5	Patient	Yes, patient reported outcomes as part of clinical trial data collection

B.6	Others Healthcare- state	Examples needed
B7	Others Social care	No??
B8	Death	Yes
B9	Census	Race/ethnicity data
B10	Existing registries	Yes???
B11	Others- state	Provision of a linkage mechanism is critical. In order to conduct analyses that require data from different datasets, a unique patient code is required to source data about an individual from the different datasets and link this into a single record.

C) Temporal needs of the data

a) detailed patient histories are required for protocol assessments and for identification of potential subjects. Retrospective data going back 10 years (as with currently available primary care based databases) are adequate.

b) currency of data: for protocol assessments, typically data that are current to within weeks will be acceptable. For screening and recruitment, up to date information would be needed. During data collection, up to date information would be required

D) Size/population of dataset required

England only
UK wide
Other combination- please state

For clinical trials research, the more the better, but complete coverage is not required. Systems that link in other parts of the UK would be preferred to an England-only system.

E) Source data verification

Yes – must have the ability to validate coding in order to comply with regulatory requirements for clinical trials research

F) Level of identification required

Fully anonymised: this is sufficient for protocol assessments during the clinical trial planning stage

Pseudonymised: it is essential that once the subject has given consent to participate in the clinical trial, he/she can be contacted (researchers at any stage during data collection, analysis or reporting could flag a suspected adverse event from the data)

Personally identifiable information: the investigator and the sponsor will need to work with this once a patient has given consent for their participation in the clinical trial.

At whatever level of identification, in order to conduct analyses that require data from different datasets, a unique patient code is required to source data about an individual from the different datasets and link this into a single record.

G) Dictionaries

Ongoing data collection should use accepted recording criteria to maximise the validity of data for protocol assessments (i.e. to minimise false positives in the trial inclusion criteria and false negatives in the trial exclusion criteria,). Coded records should be linked to structured dictionaries to allow the aggregation of individual records into broader disease classes. Where possible, a mechanism should exist to map data recorded using historical systems (e.g. Oxmis/Read) to current dictionary systems (e.g. Meddra, Snomed). However, the validity of the information relies on access to the data originally recorded by the healthcare professional. Although a mechanism to map historical data is required, all original recorded data should remain intact. These data should not be recoded prior to being made available for research.

H) Timings /data delivery/governance/pre-study

As well as providing data services (e.g. access to patient level data, standard and custom reports, expert advice for researchers), SUS must also ensure

10) Guide to the Secondary Uses Service, Description and Data Sources, released by NHS Connecting for Health on 13th November 2006

that appropriate mechanisms are in place to provide access back to sites of data collection. A secure 'third party' function will be required to allow patients/GPs to be approached for recruitment and to facilitate verification of records against the source data. Strict governance will be needed to maintain the continued anonymity of the patients (and healthcare providers) within the datasets.

The comments above are augmented with more detail on the requirements in the Appendix.

At this stage, it is understood that SUS will be able to provide access to some data sources, but much of the data needed for clinical trials research would not be included. Primary care data and laboratory data, for example, would not be available under the current proposals. From Table 2 of the most recent description of the data and services to be provided by SUS[10], some of the data would be valuable, some would already be accessible through primary care or other records, and some would not be linked.

Table 2 SUS Data Sources and their relevance to Clinical Trials Research

Data set	Relevance to Clinical Trials
National Care Records Service Data	
Personal Demographics Service (PDS)	Limited but useful data. May be more useful to source this data through primary care records
Choose and Book (CAB)	Low
Electronic Prescriptions Service (EPS)	High, but no historical data
Personal Spine Information Service (PSIS) (summary care record)	Useful but limited data. May be more useful to source this data through primary care records rather than PSIS
NHS National Collections	
Commissioning Data Set (CDS)	Some, eg diagnosis information would be useful, but would need to be linked
Mental Health Minimum Data Set (MHMDS)	
Quality and Outcomes Framework (QOF) and Quality Management and Analysis System (QMAS)	Low
Specialist Collections	
Cancer Waiting Times (CWT)	Low
Clinical Audit for Diabetes (NCASP - National Clinical Audit Support Programme)	Some useful data
Clinical Registrations (Cancer, Renal)	
Workforce Executive Information System (WEIS)	Low
Performance Management information for the Department of Health and NHS (UNIFY)	Low
Office of National Statistics	
Birth Notifications	Low
Death Notifications	High (if linked). ICD diagnosis useful
Reference Data	

Some data and some services could usefully be supplied centrally (see key issues overleaf)

5. Key issues emerging

5a. Major issues

From the protocol assessment and the simulation, 12 key issues have been identified.

1) **Data from different sources will need to be linked to provide a complete picture of the patient's health and care**

A full and integrated record of each patient's health and conditions, tests (with laboratory/pathology/imaging results), drug and non-drug treatments, together with details of the care they have received in primary, secondary and tertiary care settings is essential to generate a complete picture. This would form the basis of the material for research, though only the relevant components of this information would need to be sourced for a particular research project.

More specifically, subjects being considered for a clinical trial need to satisfy a range of criteria in order to be eligible for participation. With the full and integrated record, patients can be identified more quickly and efficiently, and fewer will need to go through expensive screening for eligibility. For subjects taking part in a clinical trial, obtaining a reliable record of concomitant medications taken during the study, including detail from the pharmacy record on the medications dispensed, will enhance our ability to address safety questions.

Linking data from different sources for a patient will require a unique code for each patient (ensuring systematic use of the NHS number would be one approach), together with investment in establishing and maintaining linkages across a range of different sources.

2) **Population coverage needs to be high**

Clinical trial planning relies on the ability to link back to subjects, typically via investigators, who may be suitable for inclusion in a study. Although protocol assessments based on a sample of the population may be suitable for estimating the prevalence of disease, with absolute numbers of expected subjects then estimated by extrapolation, without widespread coverage of the population, the numbers from a full protocol assessment will often be too low for the analysis. Sample sizes from currently available datasets are typically insufficient to highlight individual sites for recruitment. Decisions regarding the choice of investigators need to be based on knowledge of the true number of potential subjects available at a site, rather than estimates derived from population figures. These decisions can only be made with full information about the population within that site.

Currently available datasets are neither complete, nor widespread in their coverage. The coverage from the NHS CRS needs to meet both of these to facilitate efficient recruitment and to avoid bias in trial planning and other epidemiological studies. In practice, provided the population coverage is widespread and representative of the total population, achieving full population coverage will be less important than having complete patient-level data.

3) **Real-time data will often be required for clinical trial research**

Data will need to be recent for most research, and for many purposes in clinical trials research, they will need to be real-time. Where possible subjects for a trial, or other prospective research studies, need to be identified, it will be important to characterise their current health status and treatments. Without this in clinical trials, many more potential subjects need to go through expensive screening processes than would otherwise be necessary.

For data to be available real-time, there would either need to be systems in place for SUS to upload comprehensive data on a frequent basis (nightly) or some other set of systems available to build data marts from the different, up-to-date, components of the patients' records.

4) Without the addition of historical data, it will be years before much research will be possible

The data collected through the NHS CRS system will need to be supplemented with information already available to provide patient histories from Day 1. Most retrospective research, including protocol assessments for clinical trials, and most prospective research, including identification and recruitment of subjects for clinical trials, requires reference to the subject's history of their health and care.

Adding historical data to the patient's record will not necessarily require entering a substantial amount of data into an electronic system for the first time: there is a substantial amount of patient data from primary care, for example, already held electronically in existing systems.

5) Research relies on data that are quality assured

The validity of any research finding relies on the quality of the original source data. Processes and guidance are required to ensure that data recording meets threshold standards of data quality. Data should be complete, continuously collected, internally consistent and, in the case of data for clinical trials, readily available for source data verification. Currently, the Quality and Outcomes Framework (QOF) provides guidance in the required standards of recording for specific disease areas within general practice. Existing research UK databases, such as GPRD, provide recording guidelines, and continuously monitor the quality of data, before releasing these data for research.

For clinical trials research, assurance of quality is required at all stages. During 'Study Design and Planning' and during 'Initiation and Recruitment', errors in the data would typically lead to inefficiencies in the study; during 'Study Conduct' and any follow-up, errors could also lead to more serious, safety issues. There is a particularly strong need therefore for data collected as part of a study to be rigorously quality assured.

A capability to monitor and assess key elements of data quality would need to be in place for the data available to be considered 'quality assured'. This monitoring capability could also be used to provide feedback on the quality of the data maintained for the clinical management of patients.

6) Data Governance must be robust and capable of facilitating research

Data governance procedures are required to protect the anonymity of an individual patient without compromising the ability of researchers to conduct clinical studies.

Any re-use services proving additional information about particular individuals (for example, validation services) require a trusted independent party that can identify the patient of interest (or more likely this patient's GP), without revealing the patient's identity to the researcher. The separation of functions ensures that no personally identifiable information becomes available to researchers. Whereas some validation services may be able to be managed centrally, trial recruitment may rely on local groups identifying and contacting GPs and patients to invite participation in a clinical trial.

Re-use services would need to provide resources to identify the subject of interest, and pass these details to an independent body to facilitate recruitment. Additional services may be required so that further information (for long-term follow-up studies, for example) could be extracted subsequently on behalf of the researchers, without revealing the anonymised identifier linked to the individuals in the trial.

7) Access to patient level data is critical

Researchers need access to data (in general, anonymised data) at the level of the individual patient, such that all treatment and outcomes are linked to the relevant patient. Because activities for a clinical trial, such as protocol assessment and identification of potential subjects for the study, in each case reflect a unique protocol design, each activity is a

bespoke study requiring data to be analysed at the level of the patient.

For this to be feasible, the data from the proposed Secondary Uses Service (SUS) need to be made available as patient specific records (rather than reports or summary tables), requiring a repository or series of data marts covering all patients in England, and in each case linked to the details of the patient demography, health, treatments and care.

8) Adopting standard controlled terminologies will allow efficient data processing, aggregation and analysis

Regulatory agencies specify the controlled terminologies ('standards') to be used for clinical trial reporting, such as LOINC codes, SNOMED, MedDRA. Most eHR systems collect structured data that are linked to standard codes that can be mapped to the terminologies required for reporting. Some essential data are captured in unstructured text e.g., progress notes, some findings reported by sources outside the GP office. For any clinical trial work, it is essential to have the original verbatim terms and the system codes used to categorise the data. A reliable process for accurately summarising or mining unstructured data and mapping coding systems to controlled terminologies would be ideal.

This mapping can be costly and time-consuming; a re-use service would need to be able to provide complete and accurate dictionaries to assure the reliability of data interpretation. Validated tools with automated matching algorithms can relieve some of this burden.

9) Data supplied to regulatory agencies needs to be collected and analysed in a validated environment

Regulatory agencies require that all results should use validated processes & systems. It is imperative that the relevant systems comply with International Guidance and EU Directives concerning the conduct of clinical research (International Conference on Harmonisation (ICH)-Good Clinical Practice, Declaration of Helsinki, EU Directive 2001/20/EC) as well as local regulations concerning the use and storage of data. Systems supplying data for regulatory activities require specific documentation to confirm that they are fit-for-purpose, that they have been tested and functioned as expected, and have 'change-control' processes in place to account for later changes or updates.

It would be expected that documentation confirming the integrity of systems would be maintained to support the primary functions of NHS CRS. Similar documentation would also be required for other systems supporting reuse services and any additional modules specifically for electronic data capture for (e.g.) clinical trials. In addition, if a system goes down then, in the same way that a back-up system will need to be activated for emergency medical situations, a back-up system to protect the integrity of clinical trial data that continues to be collected will need to be in place.

10) Robust and secure mechanism needed to link back to investigators and subjects for clinical trials

Research that informs clinical trial planning and recruitment strategies will be little more than an intellectual exercise without some reliable and secure way to reach out to the investigators and subject populations highlighted through protocol assessment and prospective health services research.

Processes and facilities will need to be put into place for data governance for clinical trial research purposes (including, where appropriate, the need for consents or permissions, methods of anonymisation or pseudonymisation, for example) to ensure ethical research practice and data confidentiality.

11) Processes and tools for data collection for clinical trials need to be designed and implemented

The researchers' vision of 'research at the point of care and care at the point of research' dovetails neatly with the NHS migration towards paperless practice. Clinical research using dedicated electronic data capture systems requiring data entry separate from the eHR is as burdensome and outmoded as using paper CRF checklists for chart review and supplementary data collection. Methods for unobtrusively collecting electronic health care data at the point of care exist, including customised protocols and alerts designed within the system. Scenarios have been developed for electronic source data interchange and mapped to key regulatory requirements from the International Conference on Harmonisation (ICH) and the FDA[11].

An expert re-use services system would take the next step towards facilitating direct point-of-care electronic healthcare research by testing and validating one or more models for data collection. The reduction in paperwork and duplicative data entry would allow more physicians to participate in clinical research without sacrificing patient care and clinic support.

12) A capability for supplying data and associated services needs to be built

It will not be sufficient to simply supply the data. In order to make the most of the data and its potential, a research and support service will be required. These services will range from refreshing the data, documentation and related information, eg dictionaries, through to a full research capability not only for conducting studies, such as protocol assessments, but also for advising on best practice in research using the data.

To build and manage a team of developers, coding experts, programmers and qualified researchers will require investment up front, but the cost could be recouped through the provision of data and services for fees.

5b. Implications for the way forward

While there are very good data in certain parts of the country or covering certain types of medical record, there is still much to be done to extend this; to bring this together in an integrated form; and to provide the data and capabilities to capture this for the benefit of research. Our understanding from discussion with CfH staff is that some of this would indeed be available through the NHS CRS, much would not.

There is, however, a wealth of relevant experience within the UK of developing specific or local systems and capabilities to support clinical research. While we have seen some of the work being done in Scotland and Manchester, and have drawn upon the years of experience that have built the GPRD as part of our protocol assessment, we are aware that there are many other examples around the country.

From this simulation, we do not expect a solution solely based on SUS to be the best way forward, but that hybrid approaches should be explored and pursued, either a confederation of local integrated databases and capabilities, or a federation of national datasets and capabilities (or a mixture of both). In each case there would still be some things that would need to be done centrally and SUS could play that role.

11) CDISC Electronic Standard Data Interchange (eSDI) group, (2006), Leveraging the CDISC standards to facilitate the use of electronic source data within clinical trials. eSDI. http://www.cdisc.org/eSDI/index.html

6. Summary

We remain convinced that the NHS CRS has the potential to deliver significant value for the future of clinical trials research in the UK and therefore, CfH should be given a broader set of objectives to encompass realising the research applications as a core goal, rather than an add-on.

From this simulation, we do not expect a solution solely based on SUS to be the best way forward, rather hybrid approaches should be explored and pursued in order to build upon the wealth of experience and local successes that already exist around the country.

As one of the purposes of the simulation reports is to 'inform plans for full pilots to test the capacity of the infrastructure using real patient data with appropriate safeguards when this becomes feasible'[12], we suggest that two early pilots be conducted using existing data sources to provide guidance on, and highlight issues around, developing hybrid approaches:

- Merger of a primary care dataset with the HES hospital data set. This would also provide immediate value to researchers nationally
- Merger of two integrated local healthcare datasets, for example Salford and Wirral. This would provide immediate value to healthcare practitioners and researchers locally and nationally.

These exercises would also provide more details on what at a minimum needs to be done centrally under the SUS umbrella to provide the necessary national systems and capabilities, and to give a common direction for future development of the NHS CRS.

7. List of Appendices:

Appendix 1
Report from the protocol assessment of the ADOPT clinical trial (summary)

Appendix 2
Table listing the specific information about the data elements required for a 'minimum data set'

Appendix 3
Summary of the standard types of data collected during a clinical trial, organised in the way required by the FDA.

Appendix 4
Comparison of clinical trial data requirements for regulatory agencies with data types available in GPRD

(From the information in Appendix 3, this table has been drawn up to demonstrate the similarity in data required for clinical trials and data to be available through NHS CRS (using the classification of files currently available through the largest database in the UK based on primary care records, the General Practice Research Database, GPRD)).

12) Minutes of the R&D Advisory Group to Connecting for Health meeting held on 28th July 2006

Appendix 1

Protocol assessment of the ADOPT clinical trial (summary)

Introduction

A protocol assessment is a descriptive study where the proposed inclusion and exclusion criteria defining the subjects to be entered into the clinical trial are applied to a real-world patient population. The number of subjects likely to be available for recruitment is assessed, together with the likely impact of adjusting individual criteria.

The ADOPT trial, conducted in primary care, compared the effects of long-term treatment on the improvement of glycaemia control in patients with recently diagnosed type 2 diabetes mellitus. An assessment of the trial is presented here, based on data from the UK General Practice Research Database (GPRD).

The GPRD contains detailed information on diagnoses, prescribing, investigations, risk factors, outcomes and hospital referrals, together with basic demographic information for approximately 5.3 million patients from more than 370 representative general practices throughout the UK. The database is population-based and representative of the age, sex and geographic regions of the UK (Rodriguez, 1998)[13], hence results are broadly applicable to the wider UK population. The completeness and accuracy of the recording has been validated externally (Jick H 1991[14], Jick H 1992[15], Jick SS 2003[16]).

The purpose of the protocol assessment in this exercise was to identify technical, operational and governance issues that would need to be addressed if the NHS CRS is to be designed to support protocol assessments in the future, rather than to provide a full set of quantitative results and associated interpretation to guide recruitment questions.

Methods

The principal inclusion criteria for the trial required patients to be aged between 30 and 75 inclusive, to have had diabetes for 3 years or less, to be drug naïve (for glycaemia control), and with a pre-screening fasting plasma glucose (FPG) level of 126-240 mg/Dl.

To consider the impact of the ADOPT trial criteria, a fixed recruitment date of 01/01/2006 was set and the characteristics of patients registered in GPRD on this date were assessed. The date of a new (incident) diagnosis of diabetes was defined as a first ever diagnosis recorded on the GPRD for patients with at least 12 months of prior registration on the database. It was assumed that patients with existing diabetes would have at least one record relating to the disease, either when they first registered with the GP or in the following 12 months. This requirement also ensured that all incident patients had at least 12 months of medical history prior to the recruitment date, to provide sufficient information for other aspects of their medical status to be assessed.

Results

Initial assessments identified 70686 patients with type 2 diabetes from a total of 1810419 eligible patients aged between 30 and 75, registered on GPRD on 01/01/2006, and with at least one year of prior registration. This was equivalent to a population prevalence of 3.9%. With longitudinal records, it was possible to establish if these reports represented new diagnoses, or the continuation of existing disease. 23196 patients were identified with recently diagnosed diabetes (<3years) by a search of the full patient histories. With the inclusion of prescribing histories, it was possible to establish if patients had received prior treatment for glycaemia control. Overall, 22061 of 23196 recently diagnosed patients

13) Use of the UK General Practice Research Database for pharmacoepidemiology. Garcia Rodriguez LA. Perez Gutthann S. British Journal of Clinical Pharmacology (1998); 45(5):419-25.

14) Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. Jick H, Jick SS, Derby LE. British Medical Journal (1991); 302: 766-768.

15) Further validation of information recorded on a general practitioner based computerised data resource in the United Kingdom Jick H, Terris BZ, Derby LE, Jick SS. Pharmacoepidemiology and Drug Safety (1992); 1:347-349.

16) Jick SS. Kaye JA. Vasilakis-Scaramozza C. Garcia Rodriguez LA. Ruigomez A. Meier CR. Schlienger RG. Black C. Jick H. Validity of the general practice research database. Pharmacotherapy (2003); 23(5):686-9.

(95.1%) had received prior treatment (69114 of 70686 - 97.8% of all patients).

The final 'screening' criterion required patients to have laboratory data indicating a fasting plasma glucose (FPG) level of 126-240 mg/Dl. Patients were excluded if they had a recorded FPG <126 or >240 mg/Dl (patients with no record on FPG in the year prior to the index date were retained in initial analyses). In total, 62623 of 70686 diabetes patients (88.6%) were identified with a FPG (or HbA1c) lab test within the preceding year, of which 20090 (28.4%) had their last reading outside the required range (see Figure 1).

Further analyses removed patients from the pool of potential subjects based on the exclusion criteria. Exclusions included a history of angina, congestive heart failure (requiring drug treatment), anaemia, lactic acidosis, renal disease, hepatic disease, or 'any clinically significant abnormality identified on the screening'. Each of these exclusions was applied in turn, based on a review of the patients' full medical history (Figure 1). Women who were lactating, pregnant, or planning to become pregnant, were excluded based on a combination of demographics (age and sex), current prescribing (contraception), current medical status (e.g. current or recent pregnancy) and medical history (e.g. hysterectomy/ menopausal status). Patients with chronic diseases requiring periodic or intermittent treatment with oral or intravenous corticosteroids, or continuous use of inhaled corticosteroids were also excluded. For this criterion, patients were excluded based on prescribing alone, with the prescription records taken as a surrogate for the underlying disease. Final criteria excluded patients with a systolic blood pressure>180mmHg or diastolic blood pressure>110mmHg, and patients with a record of drug/alcohol abuse in the preceding six months.

It was not possible to assess 'social' criteria such as the patient's ability to sign informed consent or to read and understand dosing instructions. Additionally, the data available (at present) did not allow identification of patients who were currently involved in other clinical trials.

Discussion

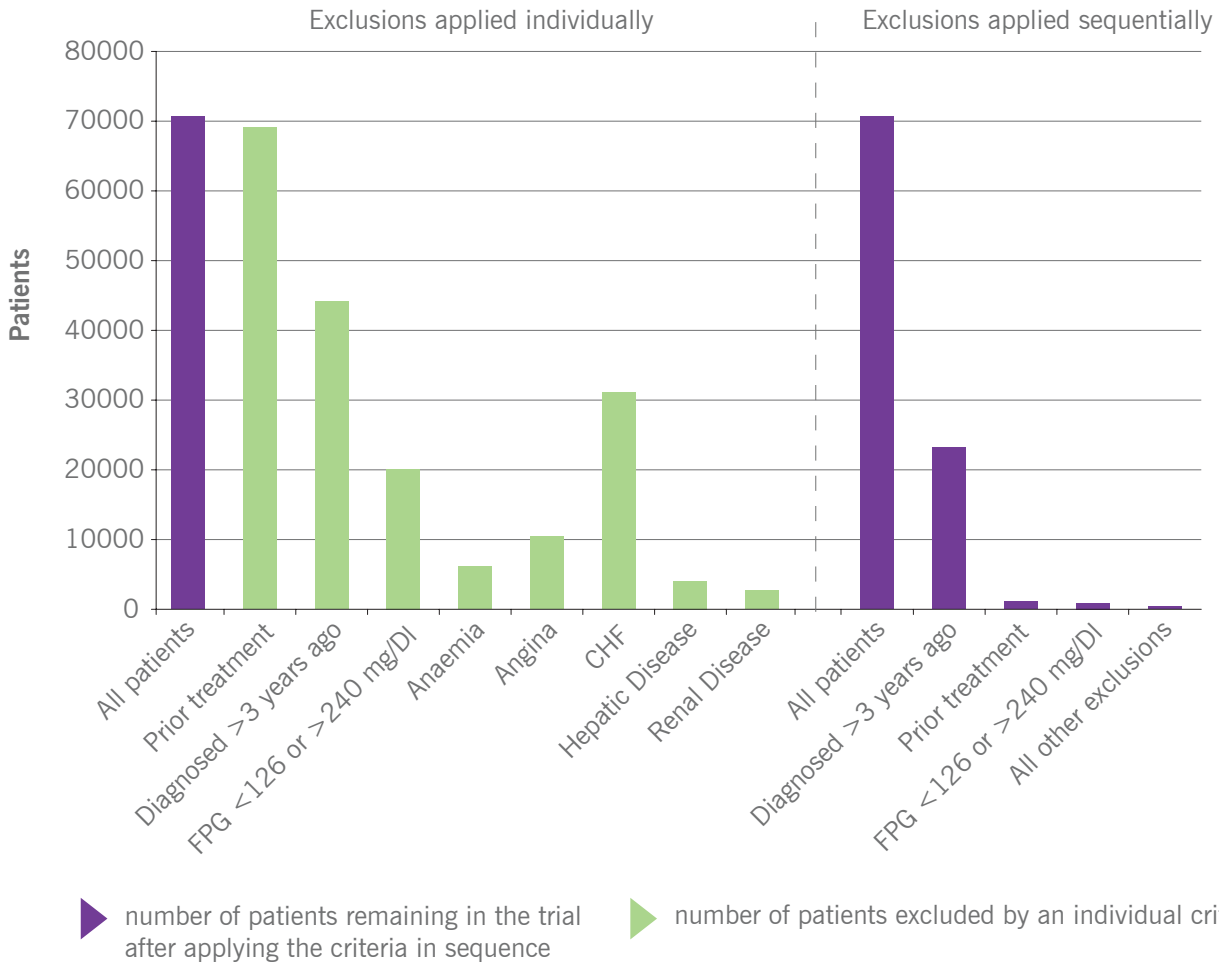
The impact of selected criteria on patient recruitment is shown in Figure 1. Only 32.8% of diabetes patients were identified as having had diabetes for less than 3 years. Additionally, 97.8% of all diabetes patients (95.1% of recently diagnosed patients) had received treatment for glycaemia control before the date of the assessment. These figures demonstrate the importance of general practice records in identifying potential subjects during the early stages of a disease and its management. In this example, it is likely that records from elsewhere in the health system, that would have identified these patients, would also have excluded the subjects from the trial (e.g. prescribing records).

Clearly, GP records are a prerequisite for studies considering diseases that are primarily managed, or diagnosed in primary care. However, they also provide a key summary of clinically significant historical information that may not be captured elsewhere in records of 'current' care. Longitudinal and historical data were required to identify incident cases, drug-naïve patients and to apply the majority of exclusion criteria. GP records may also provide the only source of information of lifestyle or risk factors. For example, both elevated blood pressure and current alcohol/drug abuse were exclusion criteria, although it is unclear if either could have been assessed without primary care records.

It is important to note, that the primary inclusion criteria required linked, patient-level data on demographics, primary care, prescribing, laboratory test results. The majority or exclusions related to significant medical events treated or diagnosed within secondary care. In primary care datasets, such as GPRD, secondary care information is limited to the summary data the GP deems clinical relevant for a patient's on-going management. The completeness and accuracy of information would be improved by access to linked data from the source, secondary care records.

In this assessment, a number of the criteria have been simplified to use currently available data. For example, the presence of hepatic disease was determined based on the presence of appropriate

Figure 1. Number of patients identified from GPRD meeting the inclusion/exclusion criteria for the ADOPT trial.



Figures are presented for selected criteria only.

diagnostic codes, rather than on laboratory values, as described in the in the ADOPT protocol. It is assumed that proxies are sufficiently robust to allow meaningful conclusions about the patient's disease status. Administrative records, such as a consultation with a specialist or a record of lab test being taken, are unlikely to allow reliable assessments of disease status. Some criteria would have been assessed more appropriately with access to other sources of data. However, the validity of the results relies on access to appropriate (and detailed) 'medical' information, rather than simple, administrative records.

Conclusions

In total, 19 of 22 criteria were assessed (to varying degrees). Overall, 15 criteria were assessed using GP 'medical' records, of which 11 required historical (as well as current) data. Three criteria used laboratory records, three used prescribing details and two included demographic details. These assessments of current medical status required access to linked, patient level records from the full range of healthcare settings, and both current and historical data. These figures could not be derived, for example, by combining simple population estimates of disease prevalence.

Appendix 2

Table listing the specific information about the data elements required for a 'minimum data set

1. **Practice.** Details of the general practice where each patient is registered.

Column name	Description
General Practice Identifier	The encrypted unique identifier given to a specific practice
Last data download	The date of the last collection for the practice
Region	NHS Region in which the practice is based

2. **Patient details.** Demographic and GP registration details for each patient

Column name	Description
Patient Identifier	The encrypted unique identifier given to a patient
General Practice Identifier	The encrypted unique identifier given to a specific practice
Year of Birth	Patient's year of birth
Month of Birth	Patient's month of birth (for those aged under 16)
Family Number	The family ID number
Sex	Sex of the patient
Currently registered	Whether the patient was currently registered
First registration date	Patient's first registration date
Registration status	The patient's current registration status
Transferred-out date	The date the patient transferred out of the practice, if relevant. NULL value for patients who have not transferred out
Transferred-out reason	The reason the patient transferred out of the practice. Includes 'Death' as an option

3. **Death Information.** ONS death certificate information

Column name	Description
Patient Identifier	The encrypted unique identifier given to a patient
Death Date	ONS recorded date of death
Place of Death	ONS recorded place of death
Deathcode1	ONS recorded cause of death 1
Deathcode2	ONS recorded cause of death 2
Deathcode3	ONS recorded cause of death 3

4. **Historical Registration.** Historical GP registration details

Column name	Description
Patient Identifier	The encrypted unique identifier given to a patient
Registration Date	The date that the patient registered with the practice.
Registration Status	The registration of the patient.
Transferred-out date	The date that the patient transferred out from the practice.
Transferred-out reason	The transferred out reason of the patient.

5. **Clinical.** Clinical diagnoses, signs and symptoms for each patient, combining records from primary, secondary and tertiary care. The setting (and site) providing the services would be identified by linking, using the Consultation Identifier, to the consultations information.

Column name	Description
Event Identifier	The unique identifier given to each event
Patient Identifier	The encrypted unique identifier given to a patient
Record Date	The date on which a specific event occurred
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Consultation Identifier	The associated consultation eid linking events at the same consultation. This links to the Event Identifier in the Consultation table.
Medical Code	The unique code for the selected medical term. This will include codes for all clinical diagnoses
Comment Identifier	The encrypted unique identifier associated with free text
Medical End Date	The date that a previous occurrence of the current clinical event resolved.

6. **Consultations.** Details of all contacts between the patient and NHS services

Column name	Description
Event Identifier	The unique identifier given to each consultation.
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Setting	Medical setting where care was delivered e.g. General practice, A&E, Dentist etc
Location Identifier	The encrypted unique identifier for the specific site providing the medical services.
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Comment Identifier	The encrypted unique identifier associated with free text

7. **Immunisation.** All immunisation records for each patient, combining records from primary, secondary and tertiary care. The setting (and site) providing the services would be identified by linking, using the Consultation Identifier, to the consultations information.

Column name	Description
Event Identifier	The unique identifier given to each event
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Consultation Identifier	The associated consultation eid linking events at the same consultation. This links to the Event Identifier in the Consultation table.
Medical Code	The unique code for the selected medical term.
Comment Identifier	The encrypted unique identifier associated with free text.
Immunisation Stage	A character indicating the stage of the immunisation given, e.g. 1, 2, B2.
Immunisation Type	The individual components of an immunisation.
Immunisation Status	The status of the immunisation, e.g. Advised, given, refusal.
Immunisation Location	The location where the immunisation was administered, e.g. In this practice.
Immunisation reason	Reasons for administering the immunisation, e.g. Routine measure.
Immunisation route	The route of administration for the immunisation, e.g. Oral, intramuscular.

8. **Referral.** Details of referrals from primary care to other healthcare services. Note: much of this information will be captured elsewhere with a fully integrated system capturing primary, secondary and tertiary care records.

Column name	Description
Event Identifier	The unique identifier given to each event
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Consultation Identifier	The associated consultation eid linking events at the same consultation. This links to the Event Identifier in the Consultation table.
Medical Code	The unique code for the selected medical term.
Comment Identifier	The encrypted unique identifier associated with free text.
National Health Service Classification	The referral speciality according to the National Health Service classification. Options are selected by the medical professional
Family Health Services Authority	The referral speciality according to the Family Health Services Authority (FHSA) classification. Options are selected by the medical professional
Referral Type	GP selected classification of the type of referral, e.g. Day case
Organisation Identifier	The encrypted unique identifier of the organisation to which the patient was referred
Organisation Type	The type of organisation to which the patient was referred
Referral Type	GP selected classification of the source of the referral e.g. GP, Self

Referral Urgency	GP selected classification of the urgency of the referral e.g. Routine, Urgent
Referral visit	GP selected category describing whether the referral event is the first visit, a follow-up etc

9. **Test.** All test records for each patient, combining records from primary, secondary and tertiary care. The setting (and site) providing the services would be identified by linking, using the Consultation Identifier, to the consultations information.

Column name	Description
Event Identifier	The unique identifier given to each event
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Consultation Identifier	The associated consultation eid linking events at the same consultation. This links to the Event Identifier in the Consultation table.
Medical Code	The unique code for the selected medical term.
Comment Identifier	The encrypted unique identifier associated with free text.
Test Type	The code associated with a specific test type.
Test text	Numerical operator (e.g. =, >, <) or string result (e.g. Abnormal).
Test Value	Numerical value of the test result.
Test Unit	The unit of measurement associated with the test result.
Lower normal Range	The start of the normal range of results for a specific test event.
Upper normal Range	The end of the normal range of results for a specific test event.

10. **Therapy.** All prescribing records for each patient, combining records from primary, secondary and tertiary care. The setting (and site) providing the services would be identified by linking, using the Consultation Identifier, to the consultations information.

Column name	Description
Event Identifier	The unique identifier given to each event
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Consultation Identifier	The associated consultation eid linking events at the same consultation. This links to the Event Identifier in the Consultation table.
Product Code	The unique code for a specific product in the product dictionary
Comment Identifier	The encrypted unique identifier associated with free text.
Prescriber source	Indicates where the prescription was issued.
Therapy Dose	Text information on product dose.
Therapy Quantity	Total quantity for the prescribed product

Therapy packs	Number of individual product packs prescribed for a specific therapy event
Pack Type	Pack size or type of the prescribed product associated with a specific therapy event
Duration of therapy	Number of treatment days prescribed for a specific therapy event

11. **Dispensing.** All dispensing records for each patient, combining records from primary, secondary and tertiary care. The setting (and site) providing the services would be identified by linking, using the Consultation Identifier, to the consultations information.

Column name	Description
Event Identifier	The unique identifier given to each event
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Record Type	The code for the category of event recorded within the recording system (e.g. diagnosis or symptom)
Consultation Identifier	The associated consultation eid linking events at the same consultation. This links to the Event Identifier in the Consultation table.
Product Code	The unique code for a specific product in the product dictionary
Comment Identifier	The encrypted unique identifier associated with free text.
Prescriber source	Indicates where the prescription was issued.
Therapy Dose	Text information on product dose.
Therapy Quantity	Total quantity for the prescribed product
Therapy packs	Number of individual product packs prescribed for a specific therapy event
Pack Type	Pack size or type of the prescribed product associated with a specific therapy event
Duration of therapy	Number of treatment days prescribed for a specific therapy event

12. **Height, Weight and BMI.** Measures of Height, Weight and BMI.

Column name	Description
Patient Identifier	The encrypted unique identifier given to a patient
Measurement date	The date the measurement was made. There may be multiple records per patient
Measurement type	May be 'height', 'weight' or 'BMI'
Measurement value	Height in meters, weight in kilograms or BMI value.

13. **Smoking status**

Column name	Description
Event Identifier	The unique identifier given to each event.
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Comment Identifier	The encrypted unique identifier which is used to link to free text.

Medical Code	The unique code for the selected medical term.
Smoking Status	The smoking status of the patient on the event date e.g. Y (Smoker), N (Never smoked), D (Ex-smoker).
Cigarettes	The number of cigarettes smoked per day.
Cigars	The number of cigars smoked per day.
Tobacco	The ounces of tobacco smoked per day.
Smoking Start date	The date the patient started smoking (can be a year, month year or a date (in any format)).
Smoking Stop date	The date the patient stopped smoking (can be a year, month year or a date (in any format)).

14. Blood Pressure

Column name	Description
Event Identifier	The unique identifier given to each event.
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Comment Identifier	The encrypted unique identifier which is used to link to free text.
Medical Code	The unique code for the selected medical term.
Systolic value	Systolic measure of blood pressure
Diastolic value	Diastolic measure of blood pressure

15. Alcohol use

Column name	Description
Event Identifier	The unique identifier given to each event.
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Comment Identifier	The encrypted unique identifier which is used to link to free text.
Medical Code	The unique code for the selected medical term.
Alcohol Status	The drinking status of the patient on the event date e.g. Y(currently drinks), N(Lifelong teetotaller), D (Ex-drinker)
Alcohol consumption	The number of alcohol units per week
Alcohol Start date	The date the patient started drinking (can be a year, month year or a date (in any format)).
Alcohol Stop date	The date the patient stopped drinking (can be a year, month year or a date (in any format)).

16. **Additional Clinical Details.** Additional information, specific to particular medical conditions (e.g. may reflect recording for QOF priorities).
These additional clinical details are not to be considered pre-requisites for the 'minimum dataset', but would provide valuable additional information for particular aspects of clinical trial research.

The examples below reflect current recording in General Practice, as captured in the GPRD database. Similar tables would be included for other specific areas, eg disease areas such as cancer or mental health, in these cases containing disease-specific information such as staging of the disease or information from patient or carer completed questionnaires.

These fields are common to all the additional clinical details files

Column name	Description
Event Identifier	The unique identifier given to each event.
Patient Identifier	The encrypted unique identifier given to a patient.
Record Date	The date on which a specific event occurred.
Comment Identifier	The encrypted unique identifier which is used to link to free text.
Medical Code	The unique code for the selected medical term.

a) Allergy and Intolerance.

Column name	Description
Reaction Medical Code	Medical code for the reaction

b) Asthma.

Column name	Description
Asthma Date	When asthma was diagnosed
Asthma status	The asthma status, e.g. Observation only, Continuous treatment, Resolved, Intermittent treatment – coded value
Asthma Risk Factors	The asthma risk factors, e.g. Not at risk, Previous history of severe attack, On 3 or more drugs, Night symptoms, Recent hospital admission, Other reason – coded value
Time off Work	The patients time off in the last 3 months
Asthma Night Cough	The night cough status, e.g. No night cough, < 1/week, 1 - 3/week, > 3/week – coded value
Asthma Inhaler Ability	The asthma inhaler ability, e.g. Good, Moderate, Poor, Not examined – coded value
Peak flow 30	The treatment plan for asthma management for a patient at 30% best peak flow, e.g. Start inhaled steroids, Double inhaled steroids, Start oral steroids, Call Doctor – coded value.
Peak flow 50	The treatment plan for asthma management for a patient at 50% best peak flow, e.g. Start inhaled steroids, Double inhaled steroids, Start oral steroids, Call Doctor – coded value.

Peak flow 80	The treatment plan for asthma management for a patient at 80% best peak flow, e.g. Start inhaled steroids, Double inhaled steroids, Start oral steroids, Call Doctor – coded value.
Steroids	Whether the patient was given oral steroids in association with their last asthma attack
Nebulised	Whether the patient was nebulised in association with their last asthma attack.
Hospital	Whether the patient was hospitalised in association with their last asthma attack.

c) Diabetes.

Column name	Description
Diabetes Program	This object holds the 'type of program' data, e.g. Hospital, Practice, Shared – coded value
Check Up Type	This object holds data on the type of check up, e.g. Acute, Annual, Routine – coded value
Left ankle jerk	This object holds the 'left ankle jerk' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Right ankle jerk	This object holds the 'right ankle jerk' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Left ankle vibration sense	This object holds the 'left ankle vibration sense' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Right ankle vibration sense	This object holds the 'right ankle vibration sense' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Left eye visual aid	This object holds the 'visual aid' data, e.g. None, Wears contact lenses, Wears spectacles – coded value
Right Eye visual aid	This object holds the 'visual aid' data, e.g. None, Wears contact lenses, Wears spectacles – coded value
Right Dorsalis pedis pulse	This object holds the Left Dorsalis pedis pulse' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Right Dorsalis pedis pulse	This object holds the 'Right Posterior tibial pulse' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Left Dorsalis pedis pulse	This object holds the 'Left Dorsalis pedis pulse' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value
Right Dorsalis pedis pulse	This object holds the 'Left Posterior tibial pulse' data, e.g. Absent, Diminished, Increased, Normal, Not examined – coded value

d) Diet.

Column name	Description
Diet Type	This object represents the type of diet of the patient.
Vegetarian	This object represents the type of vegetarian diet of the patient.

e) Exercise.

Column name	Description
Exercise taken	This object represents the type of exercise.

f) Maternity.

Column name	Description
Expected Delivery Date	This object holds data on the 'Expected Delivery Date'. This field can be entered manually or automatically calculated from the date of the last menstrual period.
Certainty of the Expected Delivery Date	This object holds data on the certainty of the Expected Delivery Date, e.g. Sure, Unsure, Very unsure – coded value
Number of births	This object holds data on the number of births
Number of miscarriages	This object holds data on the number of miscarriages
Gestation Weeks	This object holds data on number of weeks gestation related to an ante-natal booking.
Type of ante-natal booking	This object holds data on the type of ante-natal booking , e.g. GP unit, Home delivery, Hospital – coded value

g) Residence.

Column name	Description
Type of residence	Type of residence
Occupants	Lives alone

h) Sleeping Pattern.

Column name	Description
Sleep pattern	Sleep pattern
Sleep Hours	Average hours per night

i) Treatment Compliance.

Column name	Description
Treatment compliance	Treatment compliance

Finally, it would also be useful to include fields relating to participation in clinical trials to record:

- Whether an individual is currently participating in a clinical trial
- Inability to consent, linking to a contact or carer for trial purposes, as in some disease areas a significant proportion of potential subjects could be incapacitated or minor
- Whether the subject has a Health Space record (as, for example, much of the valuable screening data could have been included by the subjects themselves).

Appendix 3

Description of standard SDTM clinical trial data sets and domains required for submissions to FDA

CLASS	DOMAIN	Description	Structure
SPECIAL PURPOSE	DM	Demographics	One record per subject
INTERVENTIONS	CM	Concomitant Medications	One record per medication per subject
INTERVENTIONS	EX	Exposure	One record per constant dosing interval per subject
INTERVENTIONS	SU	Substance Use	One record per substance use per subject
EVENTS	AE	Adverse Events	One record per event per subject
EVENTS	DS	Disposition	One record per disposition status per subject
EVENTS	MH	Medical History	One record per medical condition per subject
FINDINGS	EG	ECG	One record per ECG measurement per subject
FINDINGS	IE	Eligibility Exceptions	One record per I/E criterion not met per subject
FINDINGS	LB	Labs	One record per lab test per subject
FINDINGS	PE	Physical Exam	One record per subject or body system
FINDINGS	QS	Questionnaires	One record per question per subject
FINDINGS	SC	Subject Characteristics	One record per subject characteristic
FINDINGS	VS	Vital Signs	One record per vital sign measurement per subject
SPECIAL PURPOSE	CO	Comments	One record per comment per subject
TRIAL DESIGN	TE	Trial Elements	
TRIAL DESIGN	TA	Trial Arms	
TRIAL DESIGN	TV	Trial Visits	
TRIAL DESIGN	SE	Subject Elements	
TRIAL DESIGN	SV	Subject Visits	
TRIAL DESIGN	TI	Trial Eligibility Criteria	
SPECIAL-PURPOSE RELATIONSHIP	SUPPQUAL	Supplemental Qualifiers	
SPECIAL-PURPOSE RELATIONSHIP	RELREC	Relate Records	
SPECIAL-PURPOSE RELATIONSHIP	RELDS	Relate Datasets	

Appendix 4

Comparison of clinical trial data requirements for regulatory agencies with data types available in GPRD

GPRD Filename	DOMAIN[17]	Data set examples	Type	Format[18]	RTR[19]	PRO[20]	E2[21]	E3[22]
Roster File	DM	GP or Consultant	char		Y	Y	Y	Y
Demographic File	DM	Unique Subject ID	char		Y	Y	Y	Y
	DM	Birthdate (year)	char	YYYY	Y	Y	Y	Y
	DM	Sex	char1	M,F,U	Y	Y	Y	Y
	DM	Ethnicity/Race	code		*	Y	N	Y
Life style File	VS	Weight	num (kg)		Y	Y	Y	Y
	VS	Height	num (cm)		Y	Y	Y	*
	VS	Body Mass Index	num		*	*	N	*
	PE	Stomach Girth	num (cm)		*	*	N	*
	VS	Vital signs	char		Y	Y	*	Y
	SU	Smoking status	code		*	*	N	*
	SU	Current smoking #	char1	Y/N	*	*	N	*
	SU	Alcohol consumption	num/unit		*	*	N	*
	MH	Diet	code/text		*	*	N	*
	MH	Exercise	code/text		*	*	N	*
Drug + Treatment File	EX	Substance name	code/text		Y	Y	Y	Y
	EX	Lot/batch number	code/text		N	Y	Y	Y
	EX	Start date	char	ISO8601	Y	Y	Y	Y
	EX	End date	char	ISO8601	Y	Y	Y	Y
	EX	Form	code		Y	Y	Y	Y
	EX	Strength	code		Y	Y	Y	Y
	EX	Dose	num		Y	Y	Y	Y
	EX	Dose unit	code		Y	Y	Y	Y
	EX	Route	code		Y	Y	Y	Y
	EX	Duration	num/unit		Y	Y	Y	Y
	EX	Category	code		N	Y	Y	Y
	CM/AE	Other treatments	code/text		Y	Y	Y	Y

*Y=Required N=Not Required *=Sometimes required*
#=Required for report if outcome is hospitalization or medical care or if immunization is the focus of the reported event

17) SDTM: CDISC Study Data Tabulation Model, format expected by FDA reviewers
18) Where specified, this is the format required for SDTM submission data sets
19) RTR: Retrospective studies (Clinical Trial Protocol Assessments, Trial Planning and Recruitment Models)
20) PRO: Prospective studies: Electronic Data Capture/Transfer, Extension Studies and Post-marketing Surveillance
21) E2: ICH E2 safety reporting requirements
22) E3: ICH E3 clinical study report requirements

GPRD Filename	DOMAIN[17]	Data set examples	Type	Format[18]	RTR[19]	PRO[20]	E2[21]	E3[22]
	CM/AE	Drug class	code/text		Y	Y	Y	Y
	CM/AE	Drug indication	code/text		*	Y	Y	Y
	CM/AE	Allergies, intolerances	code		Y	Y	Y	Y
Clinical File	MH/IE	Current illnesses	code		Y	Y	N	Y
	MH/IE	Past illnesses	code		Y	Y	N	Y
	MH/IE	Current symptoms	code		Y	Y	N	Y
	MH/IE	Past symptoms	code		Y	Y	N	Y
	MH/IE	Presenting details	code/text		Y	Y	N	Y
	PE/IE	Examinations	code	LOINC	Y	Y	N	Y
	PE/IE	New, ongoing issue	char1	N/O	Y	Y	N	Y
	MH/AE	Hospitalised y/n	char1	Y/N	Y	Y	#	Y
	MH/AE	Out-patient clinic	code/text		Y	Y	#	Y
	MH/AE	Accident/emergency	code/text		Y	Y	#	Y
Immunisation/ Preventative	EX/CM/AE	Immunisation batch	code		Y	Y	#	Y
	EX/CM/AE	Immunisation seq nu	code		Y	Y	#	Y
	EX/CM/AE	Immunisation refusal	code		Y	Y	#	Y
Referral File	MH/AE	Hosp admit date	char	ISO8601	Y	Y	#	Y
	MH/AE	Hosp discharge date	char	ISO8601	Y	Y	#	Y
	MH/AE	Hosp discharge diagnosis	code/text		Y	Y	#	Y
	MH/AE	Hospitalisation	code/text		Y	Y	#	Y
Test/Examination File	LB	Lab test date[7]	char	ISO8601	Y	Y	N	Y
	LB	Laboratory test	code	LOINC	Y	Y	N	Y
	LB	Lab test results	num/units		Y	Y	N	Y
	LB	Upper limit normal	num/units		Y	Y	N	Y
	LB	Lower limit normal	num/units		Y	Y	N	Y
	EG	EKG date	char	ISO8601	*	Y	N	*
	EG	EKG finding	code/text		*	Y	N	*
	EG	Evaluator/consultant	code		*	Y	N	Y
	PE/LB	Microbiology	char	ISO8601	*	Y	N	*
	PE/LB	Examinations	code/text		Y	Y	N	*
	PE/LB	Office tests	code/text		Y	Y	N	*
	QS	Questionnaire	code/text		*	*	N	*
Death File	AE	Date of death	char	ISO8601	Y	Y	Y	Y
	AE	Cause of death	code/text		Y	Y	Y	*

*Y=Required N=Not Required *=Sometimes required*
#=Required for report if outcome is hospitalization or medical care or if immunization is the focus of the reported event

GPRD Filename	DOMAIN[17]	Data set examples	Type	Format[18]	RTR[19]	PRO[20]	E2[21]	E3[22]
Consultation File	SV	Type of consultation	code/text		Y	Y	N	*
	SV	Health Professional	code/text		Y	Y	Y	*
	DS	Disposition	code/text		Y	Y	Y	Y
Other data required for RCTs	AE	Verbatim report	text		N	Y	Y	*
	AE	Derived term	code	MedDRA	N	Y	Y	*
	AE	AE Start date time	char	ISO8601	N	Y	Y	Y
	AE	AE End date time	char	ISO8601	N	Y	Y	Y
	AE	TX emergent?	char1	Y/N	N	Y	N	Y
	AE	Action Taken	code		N	Y	Y	Y
	AE	Causality	code		N	Y	Y	Y
	AE	Serious	char1	Y/N	N	Y	Y	Y
	AE	Severity/Intensity	code		N	Y	Y	Y
	AE	Relationship	text		N	Y	Y	Y
	AE	Pattern	code		N	Y	N	Y
	AE	Outcome	code		N	Y	Y	Y
	DS	Disposition verbatim	text		N	Y	Y	Y
	DS	Disposition (coded)	code		Y	Y	N	Y
	DS	Date of disposition	char	ISO8601	Y	Y	N	Y
	QS	Question name	char		N	Y	N	*
	QS	QS subcategory	char		N	Y	N	*
	QS	Finding	code		N	Y	N	N
	SC	Subject characteristic	char		N	*	N	*

*Y=Required N=Not Required *=Sometimes required*
#=Required for report if outcome is hospitalization or medical care or if immunization is the focus of the reported event

Annex A Section 3

Prospective Tracking of a Known Cohort UK Biobank

Mr Andrew Harris
Mr Steve Walker
Professor Rory Collins
Dr Tim Sprosen
UK Biobank

1 Introduction

2 *Benefits of UK Biobank*
2.1 Reliable assessment of different causes of disease
2.2 Value of Prospective Study designs

3 Current progress

4 Specific Data Requirements
4.1 Potential participants - identification and invitation data
4.2 Baseline data, colleacted by UK Biobank
Informed consent
Clinical data
4.3 Follow-up
General Approach
Hospital records
Primary care records
Coding and validation
4.4 Participant withdrawal

5 Secondary Uses Service Data

6 Key Issues

7 Summary

1. Introduction

UK Biobank is a large blood-based prospective epidemiological study which aims to build a comprehensive resource for medical researchers. The UK Biobank resource aims to include 500,000 people from all around the UK who are currently aged 40-69. This age group is being studied because it involves people at risk over the next few decades of developing a wide range of important diseases (including cancer, heart disease, stroke, diabetes, and dementia).

This document explores the potential for the supply of participant related data. UK Biobank data requirements are outlined in the following areas:

- Potential participants - Identification and invitation
- Participant registration
- Validation of baseline data
- Follow-up
- Participant withdrawal

The National Health Service (NHS)/CfH Secondary Uses Service (SUS) is designed to provide timely, pseudonymised, patient-based data and information for management and clinical purposes other than direct patient care. The data is made available through the NHS Care Records Service. These 'secondary uses' will include functions such as healthcare planning, commissioning, public health, clinical audit, benchmarking, performance improvement, research and clinical governance.

Some of the data required by UK Biobank is recorded, or planned to be, within SUS systems. Payment by Results is the most developed project within SUS and extracts are available to NHS organisations. These extracts contain information that is relevant to UK Biobank.

2. Benefits of UK Biobank

Reliable assessment of different causes of disease

Scientists have known for many years that our risks of developing different diseases are due to the complex interplay of different factors: our lifestyle and environment; our personal susceptibility (genes); and the play of chance (luck). But, despite this longstanding awareness, a clear picture of the combined effects of different factors on the risks of different diseases in different circumstances is yet to emerge. Cohorts to date have typically been characterised by small numbers of disease cases (which may yield unstable estimates due to random variations); incomplete or inadequate measures of potential risk factors (which may yield systematic under-estimates of disease associations); incomplete or inadequate measures of confounding factors (which may yield over- or under-estimates); and/or retrospective case-control designs in which the disease itself may influence risk factor levels (i.e. 'reverse causality'). Consequently, to help

assess the main causes of various chronic diseases quantitatively, there is now a strategic need to establish some large blood-based prospective epidemiological studies in a range of settings with prolonged and detailed follow-up of cause-specific morbidity and mortality.

The UK Biobank resource aims to include 500,000 people from all around the UK who are currently aged 40-69. This age group is being studied because it involves people at risk over the next few decades of developing a wide range of important diseases (including cancer, heart disease, stroke, diabetes, and dementia). The UK NHS treats the single largest group of people anywhere in the world, and keeps detailed records on all of them from birth to death. Consequently, prolonged follow-up of participants through routine medical and other health-related records will allow the identification of comparatively large numbers of individuals who develop each of a wide range of disabling and life-threatening conditions. Because UK Biobank will involve extensive baseline questionnaire and physical measures, as well as stored blood and urine samples that allow many different types of assay (e.g. genetic, proteomic, metabonomic, biochemical and haematologic), it will be a uniquely rich resource for investigating why some people develop particular diseases while others do not. This will help researchers to understand the causes of diseases better, and to find new ways to prevent and treat many different conditions.

Value of prospective study designs

A variety of study designs can be used to investigate different aspects of the relationships between different exposures and the risk of disease. These include family-based studies of genetic factors, retrospective case-control studies of particular conditions, and prospective observational studies. For the comprehensive and reliable quantification of the combined effects of lifestyle, environment, genotype and other exposures on a variety of outcomes, a prospective study has a number of advantages. As well as allowing a wide range of different conditions to be studied, exposures can be assessed prior to disease development, which avoids recall bias and allows investigation of factors that might be affected by disease processes and treatments (e.g. blood marker concentrations, blood pressure) or by an individual's response to developing some condition (e.g. weight, physical activity, diet). Prospective studies are also able to assess those conditions that cannot readily be investigated retrospectively (e.g. fatal conditions, dementia) and can include all cases of those diseases that have high case fatality rates (e.g. myocardial infarction). Moreover, it is possible to make a broader consideration of both the risks and benefits associated with a specific exposure, through the inclusion of multiple endpoints (e.g. the full health effects of smoking on a wide range of disparate diseases; or the relevance of blood pressure to different types of vascular disease). By comparison with a retrospective design, a prospective study can also provide a more straightforward source of comparable controls selected from within the same population.

By comparison with family-based or retrospective case-control studies, much larger numbers of people need to be recruited into a prospective study and careful follow-up needs to continue for many years until sufficient numbers of cases of any particular disease have developed. Hence, for studying the impact on some particular condition of factors (such as genes) that are not likely to be materially influenced by development of that condition, alternative designs may well suffice. Family-based studies are particularly valuable for identifying genes that are causally related to disease (but may over-estimate their relevance to the general population), while retrospective case-control studies are efficient for rapid accrual of large numbers of cases of some particular disease (especially at younger ages when associations may be stronger). However even in such circumstances an established, large-scale, prospective cohort provides a valuable resource for assessing the relevance of these and other factors in the general population. Moreover, as more factors are assessed and more health events accrue over time, the UK Biobank resource will become increasingly valuable (and cost-effective) to researchers for the assessment of the complex interplay between the effects of different factors (some of which may be influenced by the development of disease and so only reliably assessed in such a resource).

For all of these reasons, several large blood-based prospective cohorts have been established in recent years, and UK Biobank is intended to complement these existing resources. Studies conducted in different populations extend the range of exposures that can be considered: for example, the 500,000 person Kadoorie Study in China involves lower cholesterol levels than can be reliably studied in the UK or other developed populations; and the 150,000 person Mexico City Prospective Study involves greater levels of obesity than in the UK. Some of these studies have concentrated chiefly on assessment of certain types of exposure (e.g. diet in the 500,000 person European Prospective Investigation into Cancer and Nutrition [EPIC], which is being conducted in several European countries) and/or of certain types of outcome (e.g. cause-specific mortality and heart disease or cancer in the Kadoorie, Mexican and EPIC cohorts), and so will be particularly valuable for assessing the relevance of those particular exposures and outcomes. By contrast, UK Biobank aims to assess the relevance of a very wide range of exposures to a very wide range of health-related outcomes (i.e. not just mortality and cancer but also many other conditions that cause substantial disability). The baseline questions and measurements have been chosen carefully to allow this wide assessment to be conducted in the whole cohort, and so too have the different blood and urine samples that are being collected and stored. In addition, there is the potential for certain enhancements to be added in substantial subsets of the UK Biobank participants to allow more detailed assessment. Moreover, by embedding UK Biobank within a single National Health Service which provides the overwhelming majority of healthcare, it is intended that a very wide range of conditions can be identified and validated.

3. Current Progress

UK Biobank aims to recruit 500,000 people from all around the UK who are currently aged 40-69, and then to follow their health long-term through medical and other health-related records. Recruitment will be via centrally coordinated identification and invitation from population-based registers (such as those held by the NHS) of potentially eligible people living within a reasonable travelling distance of an assessment centre.. This central recruitment strategy will allow invitations to be targeted to enhance generalisability and to make allowance for the impact on participation rates of various factors (e.g. age, sex, ethnicity, socioeconomic status). Each assessment centre will aim to recruit as many as possible of the nearby target population during a period of about six months to one year (depending on the local population density and transport links), and will then be relocated in order to achieve recruitment across most of the UK.

When an individual arrives at the assessment they will be asked for their consent to participate (including permission to access their medical and other health-related records for long-term follow-up), and they will then move through a series of assessment stations involving questionnaires, measurements and blood/urine sampling. This baseline assessment visit takes an average of about 90 minutes, with about 14 staff required to process over 100 people daily. Staff with an appropriate mix of nursing and technical experience will be recruited and trained specifically for UK Biobank. A fully integrated clinic IT system has been developed specifically for the assessment centre visit, with each designated station having a desk top computer linked via a secure local area network to the main assessment centre server. At the end of each day, participant data and samples will be transferred securely to the UK Biobank coordinating centre. Following sample processing in the central laboratory, multiple aliquots will be stored in an automated -80°C working archive and, at a geographically distinct location, in a back-up liquid nitrogen store for security.

The recruitment process for the study is scheduled to begin in March 2007.

1) UK Biobank - 20060725 UKB NHS Data Reqs Int Pilot v3.2.doc

4. Specific Data Requirements

4.2 Potential participants - identification and invitation data

Following discussions with the Department of Health (specifically the DoH Caldicott Guardian and the Patient Information Advisory Group), it is intended that access to NHS patient registers will be approved and provided from a single national source[1]. This will avoid recurrence of the delays in invitation mailing that were experienced in the integrated pilot phase. These delays occurred as a result of the need to gain separate access through each Primary Care Trust (PCT) that manages individual patient registers. Data transfer and subsequent processing for invitation mailing will be covered by an agreement between the Department of Health (as the data controller) and UK Biobank (as the data processor) in compliance with the 1998 Data Protection Act. It will be limited to the following information on people aged 40-69:

- **Title; forename; surname; address** - allows the person to be invited to attend a UK Biobank assessment centre where their explicit consent to participate in the study will be sought
- **Gender** - allows the invitation schedule to be adjusted in response to differential participation rates by men and women so that approximately equal numbers of both sexes are recruited
- **Date of birth** - helps ensure unique identification (avoiding repeated invitations) and allows the invitation schedule to be adjusted in response to differential participation rates by age
- **Name and address of General Practitioner (GP)** - used to inform GPs that people registered with their practices are being invited to participate. The GP identifier of a potential participant will not be disclosed
- **NHS number** – helps ensure unique identification (avoiding repeated invitations) and to facilitate rapid and reliable vital status checks through central registries immediately prior to invitation (avoiding distress to the families of recently deceased people). Although the NHS number is not essential (prior to receiving consent at the assessment visit), it would help improve the reliability of the checks for duplicate entries and recent deaths that are designed to avoid unnecessary annoyance or distress. The NHS Number would also facilitate the subsequent movement of data between the NHS and UK Biobank by ensuring that a common identifier can be made available when required

UK Biobank will receive no confidential medical information on potential participants. Date of birth and the NHS number are required to verify age and for the purposes of duplicate removal respectively.

After the end of the recruitment phase, anonymised information only, will be retained on all non-participants (i.e. did not respond or declined to participate) to allow the sampling frame to be defined with respect to: gender; month and year of birth; and Super Output Area (SOA). Post-codes for home addresses will be converted to lower layer SOAs (www.statistics.gov.uk/geography/soa.asp), which cover a minimum population of 1000 people (mean 1500) and provide information about socioeconomic class. Lower layer SOAs are built from groups of Output Areas (typically 4 to 6) and constrained by the boundaries of the Standard Table wards used for the 2001 Census. Upon conversion to SOAs, post-codes for non-participants will be safely and securely destroyed. This information will allow issues about participation rates among different groups to be addressed, and help to determine extra measures to recruit hard-to-reach groups. Subsequently, comparisons in terms of various demographic factors (such as age, gender, urban/rural, socioeconomic class) may be of relevance for considering the generalisability of the recruited cohort.

4.2 Baseline Data, collected by UK Biobank

Informed consent

As part of the initial assessment that is performed by UK Biobank, informed consent is provided by participants. Participants agree to the following terms:

- I have read and understand the Information Leaflet, and have had the opportunity to ask

2) UK Biobank - Rationale, design and development of a large-scale prospective resource

questions

- I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason
- I give permission for my medical and other health-related records to be accessed by UK Biobank for research purposes
- I give permission for long-term storage and use of my blood and urine samples for research purposes
- I understand that I will not benefit financially from participation (for example, if UK Biobank leads to commercial development of a new treatment)
- I agree to take part in UK Biobank

This consent provides the basis for UK Biobank validation that this and further data may be obtained for a specific individual. Additionally, a digital signature and copy of the consent form are retained.

Clinical Data

UK Biobank aims to assess the relevance of a very wide range of exposures to a very wide range of health-related outcomes (i.e. not just mortality and cancer but also many other conditions that cause substantial disability). The baseline[2] questions and measurements have been chosen carefully to allow this wide assessment to be conducted in the whole cohort. The assessment is summarised below:

Questionnaires	Physical measurements	Samples
sociodemographic data	blood pressure & pulse rate	EDTA
smoking & alcohol exposures	weight	plasma
family history	height (standing & sitting)	serum
early life exposures	waist & hip circumference	ACD
general health & disability	bio-impedance	buffy coats
environmental exposures	hand grip strength	red cells
dietary habits	spirometry	urine
physical activity	bone densitometry	
psychological state		
cognitive function		

4.3 Follow-up

General approach

The value of the UK Biobank resource depends not only on its ability to obtain rich baseline data and samples but also on detailed follow-up of the health of participants through their medical records.

Past medical events of all kinds are required by UK Biobank from as many different complementary sources as possible (since multiple inputs on a single event may allow it to be semi-automatically validated: e.g. GP diagnosis and referral, hospital discharge diagnosis plus laboratory results of cardiac enzymes or ECG report could allow confirmation/refutation of MI). This information will also provide a full context for the participant's medical history.

Permission will be obtained at enrolment from all participants to access all of their past and future medical and other health-related records. These health records will be used to supplement information recorded at enrolment about previous medical history, family history, investigations (e.g. radiology reports, blood tests) and exposures (e.g. medication, occupational health). Most importantly, access to such records is needed to provide follow-up information related to cause specific mortality and other health events (e.g. general practice consultations; out-patient and in-patient hospital activity; cancer and other disease registries; investigations; prescribing information).

A reliable mechanism is required for continuing to keep track of an individual participant's health records during long-term follow-up. The most reliable single identifier is the NHS number in England and Wales and the Community Health number (CHNo) in Scotland. These identifying numbers are to be obtained for all potential participants prior to their invitation to attend the assessment centre. Other identifiers (such as name, date of birth, address, and general practice) will also be obtained prior to invitation, and checked during enrolment, to allow linkage to other types of health-related information (such as occupational health records). Further information will also be sought during enrolment (including mobile telephone numbers and e-mail addresses). These different identifiers will help ensure that participants are not lost during follow up, which may continue for many decades (e.g. the NHS tracing service can use the NHS number, or name and date of birth, to obtain updated GP details and address when people move).

A variety of different sources and systems will be used to ascertain death, disease occurrence and other health-related information among participants during long-term follow-up. Some of these systems have an established track record for long-term follow-up in epidemiological studies (i.e. death and cancer registries), whereas other systems have been used less widely in such circumstances (e.g. general practice and hospital activity records), although they have been successful in particular parts of the UK (e.g. Oxford Record Linkage Study; Scottish Morbidity Record). The NHS IT systems for Scotland are already sufficiently advanced to provide an electronic link to a wide range of relevant medical records, and a substantial effort is now taking place to establish similar systems for the NHS in England and Wales. Linkage of participants within some of these systems will be initiated during the recruitment phase, but linkage to other systems will await further evolution of the central NHS IT systems. In either case, however, information will be sought from the relevant system about the participant's health from the time of their enrolment in UK Biobank and, where appropriate, from the period before recruitment (e.g. supplementing self-reported past medical history). The rest of this section describes the current and likely future availability of different types of health-related information from these different sources and systems.

Hospital records

The UK Biobank data repository needs to include information about health events and activities that are experienced by participants when they attend hospitals. While the initial referral and other information about hospital activity is likely to be recorded within the primary care record, it is important that this should be supplemented by, and validated against, the information that can be derived from the hospital systems.

The Scottish Morbidity Record (SMR) has been collecting data on all admissions to all Scottish

NHS hospitals since 1980, and these data are routinely collated by the Information and Statistics Division (ISD) of the Common Services Agency. UK Biobank's Regional Collaborating Centre for Scotland has access to methodology developed and implemented for the specific purpose of automatic retrieval of such information (e.g. the GENIE software application used successfully in the context of the national diabetes computing system). This software can be programmed to update all changes in health status for particular individuals on a daily, weekly or monthly basis by attaching an electronic flag to their CHNo in the electronic systems that hold the relevant health care information. Consequently, with the permission of the NHS Privacy Advisory Committee, UK Biobank will be able to extract hospital admission data for Scotland (and the same structures will also allow retrieval of primary care records, prescribing information, and maternity, cancer and death data).

In the medium to long term, developments in the new National Care Record Service will also allow hospital activity data for England and Wales to be retrieved from a central source. Such information is already collected at a national level for other purposes: that is, the Department of Health's Hospital Episode Statistics (HES). HES is the national statistical data warehouse for England of the care provided by NHS hospitals and for NHS hospital patients treated elsewhere. It is the data source for a wide range of healthcare analysis for the NHS, Government, and many other organisations and individuals. Data held in HES are derived from the NHS-wide Clearing Service that provides the mechanism by which HES data are transferred from individual hospital trusts' clinical systems. For each financial year, there are approximately 12 million records (episodes of care) in the HES database, which represent all NHS-funded admissions for patient care, and private care within NHS hospitals in England. (Data are not included, however, on private healthcare, activity in Accident and Emergency departments, or drugs used during the hospital episode.)

For each episode of care, HES includes information about:

- Patient identifiers (including NHS number);
- In-patient, day case and out-patient episodes (out-patient data became mandatory in October 2001 and the mental health minimum dataset mandatory from April 2003), maternity records and psychiatric census;
- Administrative details (e.g. admission and discharge date) and the organisation providing the treatment;
- Clinical information relating to diagnoses (ICD10 codes) and procedures (OPCS4 codes).

As with the SMR in Scotland, HES retain historical data that can allow UK Biobank to supplement, and validate, the information obtained at enrolment about participants' past medical history. For example, cross-referencing of validated outcomes from regular clinic (and GP) follow-up showed a very high concordance (>90%) in the Heart Protection Study with retrospective review of computerised hospital records.

Privacy of the individual is one of the basic principles behind the whole HES and SMR ethos. There are well described processes by which organisations can apply to receive this information, which is supplied as responses to specific query criteria and extracts from the core dataset. The nature of UK Biobank's request will entail special service agreements since the provision of clinical information in respect of identifiable patients is outside the normal areas of information provision to third parties. With respect to HES, SD2HES has obtained the agreement of the Security and Confidentiality Advisory Group to allow access to raw codes in specific circumstances; and, in Scotland, access to SMR data has previously been provided for such studies with the agreement of the NHS Privacy Advisory Committee. In both cases, the provision of these data to UK Biobank should be acceptable since all participants will have given signed consent at enrolment for extraction of their individual hospital records and other health-related information. It is intended, therefore, that follow-up of hospital activity through HES and SMR be initiated during the recruitment phase of UK Biobank.

Primary care records

In Scotland, as discussed above, an individual's CHI

number can already be used to link to a wide range of health-related information, including primary care, clinical and prescribing databases (e.g. GPASS in 85% of practices) going back to 1984, and systems have been developed for its automatic retrieval. Consequently, after obtaining permission from the NHS Privacy Advisory Committee and other relevant groups, it should be relatively straightforward for UK Biobank to extract general practice data for Scotland.

In England and Wales, there are numerous projects (e.g. Q-Research, EPIC/THIN and GPRD) that work directly with general practices and their current clinical system suppliers to retrieve practice data, but these do not provide national coverage. Instead, it will be more efficient to wait for the introduction of some of the infrastructure and applications that will be provided by the Connecting for Health (CfH) programme before national follow-up of primary care information for UK Biobank is started. The two key elements of CfH are the NHS Care Records Service and the Secondary Uses Service. The NHS Care Records Service will, in summary, contain the following components:

Organisational records: The electronic equivalent of detailed paper records entered by clinicians and support staff to record and plan patient care within that organisation

Detailed care records: Where organisations share the same electronic records architecture within defined geographical areas, organisational records will be shared (within the constraints of access controls)

Pathways of care and care plans: When patients have complex or chronic care needs, 'pathways of care' will indicate the local care that is normally to be delivered (with multiple pathways of care applicable to those with co-morbidity). For each patient, a single shared care plan will be derived from their separate pathways of care. The care plan will contain key relevant past events for the patient (e.g. their blood pressure measurements, by whom and when) and their planned care (e.g. who is responsible for their blood pressure monitoring, when it will next be measured, and by whom). These pathways of care and care plans will be shared by all those caring for the patient.

Summary Care Record: This will contain contributions from the general practice longitudinal record, hospital discharge and out-patient summaries, pathology and imaging results and, in time, care by others (such as social care). The Summary Care Record will be widely available to appropriate health professionals through the Personal Spine Information Service.

UK Biobank should be able to access the data in the Summary Care Record, from the pathways and journeys of care and, in some situations, from the organisational and detailed care records.

The other key programme in CfH is the Secondary Uses Service (SUS), which aims to provide 'timely, pseudonymised patient-based data and information for purposes other than direct clinical care... [including] research'. SUS will access data from all sectors of the health service and social care, including general practice, community teams, secondary care hospitals, tertiary care and private providers supplying the NHS. It will have access to the data within the NHS Care Records Service and will be able to link it to external sources, such as registration of deaths, census data and health service organisational boundaries. Certain types of health data will not be available through SUS, including care from private providers, over-the-counter and complementary therapies, self care and care delivered overseas. But, for care delivered within the NHS in England, the geographic and organisational coverage of SUS data should be close to 100%. It has been confirmed with CfH that it will be possible to use the NHS number to track and extract clinical data from SUS for research participants who have given their consent (as in UK Biobank).

In order to ensure a robust and complete data set, UK Biobank will work with SMR in Scotland, with CfH in England and Informing Healthcare in Wales to specify how access to routine clinical data for participants can be achieved (and, to that end, UK Biobank's Chief Information Officer is working on secondment within DH England, thereby enabling more direct discussions with the relevant parties). This will involve the initial specification of a historical and continuing dataset, with options for obtaining additional data from time to time to meet

the specific needs of particular areas of research. During the recruitment phase, UK Biobank is likely to be able to initiate follow-up through primary care records during enrolment in Scotland and, at least, to have established and piloted the systems for such follow-up in England and Wales.

Coding and validation

It will require several years of follow-up in UK Biobank before enough participants have developed any particular condition for reliable assessment of the main determinants of the condition. The initial recruitment phase and early years of follow-up will allow the careful development and piloting of systems for accessing and validating data from a variety of different systems. Consequently, by the time sufficient numbers of events have occurred among the participants, UK Biobank will have validated data on a wide range of health outcomes that are sufficiently reliable and complete for the purposes of most research (and that can be readily supplemented in particular ways when required for specific purposes).

Currently, the accuracy and completeness of the data available through healthcare records systems is variable, and one of the principal aims of CfH is to improve data standards and consistency. Most is known about the quality of general practice data, where early adoption of computerised systems has resulted in data quality that is often higher than in other sectors. Almost all general practices in the UK are already computerised, and up to two thirds are now using their clinical computer as the only means of recording clinical care (including encounters, diagnoses, prescriptions, etc) Moreover, the Quality and Outcomes Framework of the new General Medical Services contract for general practices has stimulated efforts to improve accuracy and completeness. Although UK Biobank may need to access some free-text entries in order to establish the exact nature of a healthcare event or decision, it will primarily use the capture and analysis of codes. Experience with Read Codes shows some variability in their use, but further education and training should help to ensure the effective implementation of Snomed Codes. More problematic is the exact meaning of certain terms: for example, while there are internationally agreed diagnostic criteria for myocardial infarction (and the patient's record is likely to include evidence that those criteria have been met: see below), no such criteria are routinely applied to post-natal depression. Moreover, clinicians are skilled at interpreting such diagnoses in their historical context (willingness to make diagnoses and use certain labels changes with time). They are also skilled at interpreting such diagnoses according to the background of the person generating the entry (e.g. different weights may be given to a label of postnatal depression that is applied by a consultant psychiatrist, obstetrician, GP or community midwife). As the health services become more reliant on electronic health records, they are shared more widely and such deficiencies become more evident. For example, analyses through SUS have revealed variations in the quality of data recording, which educational initiatives (such as PRIMIS+) are now working to rectify.

For UK Biobank, three clinical research staff have been allocated to develop and implement procedures for identification and cross-validation of outcomes from different healthcare sources. It will be important to start the process of identification and validation of health outcomes during the latter part of the recruitment phase. This will allow the coverage of healthcare outcomes to become comprehensive during the subsequent five-year period when the resource starts to become sufficiently mature for informative case-control studies of the commoner conditions (such as heart disease). As multiple sources of information about health events (e.g. primary care, hospital activity, investigations, prescriptions) become available to UK Biobank, it will be possible to build a range of semi-automatic systems for the confirmation or refutation of a wide range of outcomes that should suffice for many research purposes. For example, myocardial infarction identified from the primary care record might then be supported by a confirmatory hospital discharge record and/or by an electrocardiogram or laboratory report consistent with myocardial infarction (or, alternatively, refuted by a discharge record or investigations more consistent with, say, unstable angina). Similarly, cancer registry data may not only be confirmed but also made more specific by linking them to relevant laboratory systems (e.g. histology). These approaches will build on research

that is currently being supported through the MRC's e-science program (such as the VOTES project, which involves UK Biobank's RCCs). Even where such automated systems are not able to provide sufficiently specific information about the type of health outcome (at least in the short-term before all relevant records can be accessed); they should be able to identify a suitably limited group of individuals for whom particular information needs to be sought.

Follow-up data will be appended to the UK Biobank core data repository, and linked to pre-existing data (such as assessment visit records) primarily through indirect linkage using the participant's NHS number (validated by reference to other information, such as name and address). Some datasets may not include the NHS number, which will necessitate an auditable comparison of supplied data with other identifying data for participants (e.g. name, address and date of birth). Data that are to be included within the repository will initially be transformed into a standards based format, keeping coding structures, values and textual data in their original form in order to ensure an audit trail back to the source data. Imported data will then provide the basis for ongoing clinical validation and cross-referencing with any previously supplied data residing within the core systems. If appropriate, data may be re-coded (e.g. by conversion to a standardised coding system) or summarised to aid high-level search and querying processes which will provide more consistent information sets for subsequent data-mining and other research activities. All clinical cross-referencing and re-coding work will be performed under secure conditions, without direct reference to information that identifies a participant (such as name and address). Since initial tests on primary care data have shown high variability in the quality of coded data, it is highly likely that any automated processes will require auditable human validation and sign-off before being included within the core repository and made available for research.

4.4 Participant Withdrawal

Participants will be advised at enrolment that they have the right to withdraw at any time without having to give an explanation and without penalty. This is essential to preserve and demonstrate the voluntary nature of participation. Should participants become incapacitated or die, they would no longer be able to withdraw themselves.

During enrolment, UK Biobank will provide information to participants about the options for withdrawal:

- **'No further contact':** UK Biobank would no longer contact the participant directly, but would still have their permission to use information and samples provided previously and to obtain further information from their health relevant records.
- **'No further access':** UK Biobank would no longer contact the participant or obtain information from their health-relevant records in the future, but would still have their permission to use the information and samples provided previously.
- **'No further use':** In addition to no longer contacting the participant or obtaining further information, UK Biobank would aim to destroy all of their information and samples collected previously (although the participant would be told that it may not be possible to trace all distributed sample remnants for destruction). Only their signed consent and withdrawal would be kept as a record of their wishes. Such a withdrawal would prevent information about them from contributing to further analyses, but it would not be feasible to remove their data from analyses that had already been done.

If, having discussed their concerns and options, a participant decides to withdraw, then UK Biobank would seek written confirmation of the level of withdrawal from the participant or their legal representative. UK Biobank will need to retain some minimal personal data for a number of reasons, which include: ensuring that participants who have withdrawn are not re-contacted; and assessing the determinants of withdrawal and any impact on research findings. Participants who withdraw will be assured that this administrative record will not be part of the main database that is available to others.

Despite UK Biobank's efforts to stay in touch with participants, it may well lose contact with some as they relocate, emigrate, or do not respond to

communications. Where a participant has not actively withdrawn, UK Biobank will continue to use the samples and data and maintain linkages, although it will not be able to update some data (e.g. those collected by repeat questionnaire).

When a previously registered participant withdraws from UK Biobank, UK Biobank will inform any other data supplier that holds information regarding an individual's participation in UK Biobank, so that further information is not received either automatically (such as in the case of periodic record updates) or in the event of further erroneous requests for data from UK Biobank.

5. Secondary Uses Service Data

The Secondary Uses Service (SUS)[3] is designed to provide timely, pseudonymised, patient-based data and information for management and clinical purposes other than direct patient care. The data is made available through the NHS Care Records Service. These 'secondary uses' will include functions such as healthcare planning, commissioning, public health, clinical audit, benchmarking, performance improvement, research and clinical governance.

SUS will establish a single, secure data environment for the whole NHS. Coupled with effective analysis tools, this will enable a wide range of users to carry out comprehensive analysis in a consistent and effective way.

Information is pseudonymised (protecting patient confidentiality by allocating it a consistent 'pseudonym') and access will be governed through rigorous data access controls to protect patient confidentiality.

The first priority for SUS is the implementation of Payment by Results (PbR). The aim of PbR is to provide a transparent, rules-based system for paying NHS Trusts for the activity they undertake using a national tariff.

The Payment by Results (PbR)[4] extracts contain identifiable patient episode data for Admitted Patient Care Episodes, Outpatients and A&E. Records are categorised into specialty code, Healthcare Resource Groups and in some cases an ICD or Read diagnosis code. This data-set has the potential to provide notification of hospital based events that can be used to validate some UK Biobank baseline data and provide some follow-up data.

Over time, SUS will be enriched by other data sources including cancer waiting times, clinical audit information and central returns; as well as other non-patient-based sources, such as national statistics and workforce data. As a result, with future releases, the number of uses for SUS data will grow.

3) CfH - Guide to the Secondary Uses Service (SUS) Description and Data Sources
4) specification_of_pbr_extractsfebruary06

Table 1 SUS Data Sources and their relevance to UK Biobank

Data set	Relevance to UK Biobank	Availability
National Care Record Service (NCRS) Data		
Personal Demographics Service (PDS)	Possibly of use for matching participants. Records from 1991	
Choose and Book (CAB)	Low, includes clinical assessment data	2007
Electronic Prescriptions Service (EPS)	High	TBD (to be defined)
Personal Spine Information Service (PSIS) (summary care record)	High	TBD
NHS National Collections		
Commissioning Data Set (CDS)	High. All records from 1997. Includes HRG, ICD and/or Read diagnosis	Payment by results extracts available?
Mental Health Minimum Data Set (MHMDS)	Medium	2007
Quality and Outcomes Framework (QOF) and Quality Management and Analysis System (QMAS)	Low	TBD
Specialist Collections		
Cancer Waiting Times (CWT)	High. Includes ICD diagnosis	2007
Clinical Audit for Diabetes (NCASP - National Clinical Audit Support Programme)	High. Includes detailed diagnosis	2007 Cancer – 2007
Clinical Registrations (Cancer, Renal)	High. Includes ICD diagnosis	Cancer – 2007 Renal – TBD
Workforce Executive Information System (WEIS)	Low	
Performance Management information for the Department of Health and NHS (UNIFY)	Low	
Office of National Statistics		
Birth Notifications	Low	
Death Notifications	High. Contains ICD diagnosis	TBD
Reference Data	High	2007

Table 2 Payment by Results extracts and their relevant to UK Biobank

Extract Field Group	Extract			UK Biobank	
	Admissions	**Outpatients**	**A&E**	**Relevance**	**Priority***
Record ID	Yes	Yes	Yes	Low	3
Spell ID	Yes	Yes	Yes	Low	3
Generated Record ID	Yes	Yes	Yes	Low	3
Record Type	Yes	Yes	Yes	Medium	6
Future use (Pseudonymisation Status)	Yes	Yes	Yes	Medium	6
Patient Identifiers	Yes	Yes	Yes	High	9
Birth and age	Yes	Yes	Yes	Medium	6
Other person data	Yes	Yes	Yes	Medium	6
Spell, Admission & discharge	Yes	No	No	Low	1
Episode in spell	Yes	No	No	Low	1
no / flags	Yes	No	No	Low	1
Additional spell data	Yes	No	No	Low	1
Episode dates	Yes	No	No	High	3
Attendance Characteristics	No	Yes	Yes	Low	2
Contract Identifiers	Yes	Yes	Yes	Low	3
Provider & Commissioner	Yes	Yes	Yes	Low	3
Consultant & specialty	Yes	Yes	Yes	Low	3
GMP & referrer	Yes	Yes	Yes	Low	3
Wait & stay	Yes	No	No	Low	1
Episode level	Yes	No	No	High	3
PbR Episode level data	Yes	No	No	Low	1
Appointment Request	No	Yes	No	Low	1
Freeze and spell status	No	Yes	Yes	Low	2
HRG	No	Yes	Yes	High	6
Diagnoses block	Yes	Yes	Yes	High	9
Investigation	No	No	Yes	Medium	2
Procedures block	Yes	Yes	Yes	High	9
Augmented care	Yes	No	No	Medium	2
Tariff	No	Yes	Yes	Low	2
Interchange	No	Yes	Yes	Low	2
Query	No	Yes	Yes	Low	2

** Priority has been derived by multiplying the number of extracts that include a data item by the relevant weighting of that data item (high=3, medium=2, low=1)*

6 Key Issues

There was commonality with issues raised by other simulations. Issues of particular importance to UK Biobank are listed below:

1) **Data from different sources will need to be linked to provide a complete picture of the patient's health and care**
A full and integrated record of each patient's health and conditions, tests (with results), drug and non-drug treatments, together with details of the care they have received in primary, secondary and tertiary care settings is essential to generate a complete picture.

   Linking data from different sources for a patient will require a unique code for each patient, and investment in establishing and maintaining linkages across a range of different sources.

2) **Population coverage needs to be comprehensive**
Although UK Biobank will only require data for 500,000 individuals, these people will have a geographical spread across the whole of England, Scotland and Wales. This means that only comprehensive coverage will enable the supply of data relating to all participants

3) **Real-time data will often be required for clinical trial research**
For data to be available real-time, there would need to be systems in place for SUS to upload comprehensive data on frequent basis (e.g. nightly) or some other set of systems to build data marts from the different components of the patients' records. This is seen as a lower priority for UK Biobank, although data volumes may necessitate frequent uploads.

4) **Without the addition of historical data, it will be years before much research is possible**
The data collected through the NHS CRS will need to be supplemented with information already available in order to provide patient histories from as early a point in time as possible. Most retrospective research, including protocol assessments for clinical trials, and most prospective research, including identification and recruitment of subjects for clinical trials, requires reference to the subject's history of their health and care. UK Biobank requires both retrospective and prospective information pertaining to its participants.

   Adding historical data to the patient's record will not necessarily require entering a substantial amount of data into an electronic system for the first time: there is a substantial amount of patient data from primary care, for example, already held electronically in existing systems.

5) **Research relies on data that are quality assured**
The validity of any research finding relies on the quality of the original source data. Processes and guidance are required to ensure that data recording meets threshold standards of data quality. Data should be complete, continuously collected, internally consistent and readily available for source data verification. Currently, the Quality and Outcomes Framework (QOF) provides guidance in the required standards of recording for specific disease areas within general practice. Existing research UK databases, such as GPRD, provide recording guidelines, and continuously monitor the quality of data, before releasing these data for research.

   A re-use service would need the capability to monitor and assess key elements of data quality for these data to be considered 'quality assured'. Such monitoring could also be used to provide feedback on the quality of the data maintained for the clinical management of the patient.

5) http://www.ukbiobank.ac.uk/ethics/efg.php

6) **Data Governance must be robust and capable of facilitating research**

The key ethics and governance principles relating to UK Biobank are laid out in the Ethics & Governance Framework[5]. This was first prepared by the project funders (the Medical Research Council, Wellcome Trust and Department of Health) with the advice of an Interim Advisory Group on Ethics and Governance (IAG), chaired by Dr. William Lowrance (Geneva) and with members expert in research ethics, philosophy, law, science, social science, and consumer representation. The Group's deliberations were informed by an ethics consultation workshop in April 2002 and general consultation during 2003 on an earlier draft of the EGF with a wide-ranging group of experts and stakeholders, including members of the public, special interest groups and health-care professionals. The present draft of the EGF has been modified in the light of the developing plans for recruitment and follow-up, although it is still undergoing review by UK Biobank with advice from the independent Ethics & Governance Council and funders prior to adoption. The participant materials have been developed with the advice of the EGC and in accordance with the key principles in the EGF.

UK Biobank plans to provide identifying information such as name, address, date of birth, NHS Number (where available) and a relationship specific pseudo-identifier so that data providers are able to match and register the individual's participation in UK Biobank. The pseudo-identifier may be generated and returned by data providers as part of their confirmation. This shared and pseudonymised identifier would remove the need for subsequent use of identifying information such as the NHS Number.

UK Biobank will, in turn, be providing data (pseudonymised and anonymised) to other research projects and clinical trials for re-use. Data governance procedures are required to protect the anonymity of individual patient without compromising the ability of researchers to conduct clinical studies.

Any re-use services proving additional information about particular individuals (for example, validation services) require a trusted third party who can identify the patient of interest (or more likely this patient's GP), without revealing their identity to the researcher. The separation of functions ensures that no personally identifiable information becomes available to researchers. Whereas some validation services may be able to be managed centrally, trial recruitment may rely on local groups identifying and contacting GPs and patients to invite participation in a clinical trial.

Re-use services would need to provide resources to identify the subject of interest. Additional services may be required so that further information (for long-term follow-up studies, for example) could be extracted subsequently.

Processes and facilities will need to be put into place for data governance for research purposes (including, where appropriate, the need for consents or permissions, methods of anonymisation or pseudonymisation) to ensure ethical research practice and data confidentiality.

More specifically, UK Biobank is keen to know what steps would be required in order to:

- Register & manage a 'legitimate relationship' with specific patients (cohort)
- Gain access to patient level data via SUS
- Gain access to the summary care record
- Gain access to detailed care records
- Gain access to the sealed envelope
- Connect directly to PSIS/spine/LSPs

7) **Access to patient level data is critical**

UK Biobank needs access to data at the level of the individual patient, such that all treatment and outcomes are linked to the relevant patient.

Data needs to be made available as patient-specific records rather than reports or summary tables.

Informal trials by UK Biobank on primary care data have shown the quality of coding to be highly variable. For example, 100 sample GP to GP extracts, including some 3,000 clinical recordings, have shown the vast majority (>95%) of clinical observations to be coded as 'General comment' (9B04) with no relation to the information held within the text/annotation. Regardless of the level of quality assurance associated with supplied data, it is likely that UK Biobank will need to internally validate records and perform further analysis within and between data sets in order to extract maximum value from the data.

8) **Adopting standard controlled terminologies will allow efficient data processing, aggregation and analysis**

Most eHR systems collect structured data that are linked to standard codes that can be mapped to the terminologies required for reporting. Some essential data are captured in unstructured text (e.g., progress notes, some findings reported by sources outside the GP office. For some research, it is essential to have the original verbatim terms and the system codes used to categorise the data. A reliable process for accurately summarising or mining unstructured data and mapping coding systems to controlled terminologies would be ideal.

This mapping can be costly and time-consuming; a re-use service would need to be able to provide complete and accurate dictionaries to assure the reliability of data interpretation. Validated tools with automated matching algorithms can relieve some of this burden.

9) **Data supplied to regulatory agencies needs to be collected and analysed in a validated environment**

10) **Robust and secure mechanism needed to link back to investigators and subjects for clinical trials**

11) **Processes and tools for data collection for clinical trials need to be designed and implemented**

The researchers' vision of 'research at the point of care and care at the point of research' dovetails neatly with the NHS migration towards paperless practice. Clinical research using dedicated electronic data capture systems requiring data entry separate from the eHR is as burdensome and outmoded as using paper CRF checklists for chart review and supplementary data collection. Methods for unobtrusively collecting electronic healthcare data at the point of care exist, for example customised protocols and alerts designed within the system. Scenarios have been developed for electronic source data interchange and mapped to key regulatory requirements from the International Conference on Harmonisation (ICH) and the FDA.[6]

An expert re-use services system would take the next step towards facilitating direct point-of-care electronic healthcare research by testing and validating one or more models for data collection. The reduction in paperwork and duplicative data entry would allow more physicians to participate in clinical research without sacrificing patient care and clinic support.

12) **A capability for supplying data and associated services needs to be built**

It will not be sufficient to simply supply the data: in order to make the most of the data and its potential, a research and support service will be required. These services will range from refreshing the data, documentation and related information, e.g. dictionaries, through to a full research capability. The services will not only be used for conducting studies, such as protocol assessments and cohort management, but also

6) CDISC Electronic Standard Data Interchange (eSDI) group, (2006), Leveraging the CDISC standards to facilitate the use of electronic source data within clinical trials. eSDI. http://www.cdisc.org/eSDI/index.html

for advising on best practice in research using the data.

UK Biobank Additional Issues

13) **Common governance terminologies and methodologies should be formally agreed**
for example: agreed definitions of the terms anonymised, reversibly anonymised, pseudonymised and agreed policies and processes for anonymisation, pseudonymisation and statistical disclosure protection.

The ONS has provided an excellent starting point for an agreed set of definitions in the Review of the Dissemination of Health Statistics - Confidentiality Guidance[7].

14) **Clarification on mechanisms for identifying UK Biobank Participants to CfH/SUS**
UK Biobank is primarily interested in patient level data for a specific cohort of up to 500,000 people in England, Scotland and Wales.

In order to facilitate and control the provision of patient level data for this cohort, management of the cohort and its relationships to NHS records is required. Systems supplying such patient level data should be configured to include cohort management tools that include such functionality as registration/de-registration of cohort members, pseudo-identifier management and data extract tools.

15) **Clarification on access to further data sets**
These include the summary care record, detailed care record and prescribing data.

16) **Sealed envelope**
It is essential to ensure that the UK Biobank consent suffices (i.e. is sufficiently explicit) to allow access to 'sealed' information in the GP record. This is seen as a high priority because any changes to the consent statements would likely need to also be approved by MREC/PIAG before the start of recruitment.

17) **Invitations**
UK Biobank requires names and addresses of individuals in specific areas (surrounding assessment centre locations) that are aged between 40 and 69. NHS Number is also required in order to perform ONS checks and to facilitate access to NHS data in the future. A centralised source is being sought. There is potential for SUS to provide this data either as a single national data feed or as a set of assessment centre oriented feeds over the life of the UK Biobank recruitment phase.

7. Summary

UK Biobank requires highly detailed data from many areas of NHS activity. High-level simulations show that the Secondary Uses Service is producing extracts that are of some relevance to UK Biobank, although they are currently limited to hospital based data sets (CDS/PbR), are only available to a subset of organisations operating within the NHS and are likely to supersede the HES Extract. Use of this data by UK Biobank is primarily dependant upon implementation of the necessary governance arrangements and completion of the necessary technical implementation by SUS.

There are plans to make further data sets available in the future – some of these, for example related to Cancer and Diabetes histories, are also of direct relevance to UK Biobank. Access to primary care related data is likely to be initially only available using the summary care record. UK Biobank is currently unclear as to both when and how it would gain access to the required NHS CRS related data-sets for its participants.

Sources:
1: **ONS: Dissemination of Health Statistics: Confidentiality Guidance: Glossary** (http://www.statistics.gov.uk/about/Consultations/downloads/Health_Stats/Health_Stats_4_Glossary.pdf)

7) http://www.statistics.gov.uk/about/Consultations/disclosure.asp

Annex A Section 4

UKCRC Connecting for Health Research Simulation Report: Observational Epidemiology

Professor Carol Dezateux
Professor Catherine Peckham
MRC Centre of Epidemiology for Child Health
Institute of Child Health, University College London

1 Introduction
1.1 Original Brief and Scope
1.2 Specific Objectives
1.3 Summary of Methods Used

2 Opportunities for Record Linkage
2.1 The Role of Observational Epidemiology
2.2 Benefits of Record Linkage
2.3 Assisted Reproductive Technologies

3 Current Situation
3.1 Current Data sources in England
3.2 Current UK data sources for ART
3.3 NHS CRS and SUS

4 Proposed Solutions

References

1 INTRODUCTION

1.1 Original Brief and Scope

We were commissioned to undertake a simulation relevant to observational epidemiological research, based on retrospective analysis of data using case-control and cohort study design, in order to:

1) inform any future development of the NHS Care Record Service

2) highlight regulatory and governance issues that need addressing

3) inform plans for any further simulations and full pilots to test the capacity of the infrastructure using real patient data with appropriate safeguards when this becomes feasible

4) where possible, to examine options for interoperability on a UK-wide basis by considering approaches in England, Scotland, Wales and Northern Ireland to routine data collection and use

1.2 Specific Objectives

1) to assess the extent to which the current arrangements proposed within Connecting for Health (CfH) support the conduct of high quality observational epidemiological research studies based on retrospective analysis of routine data

2) to use simulated research case studies based on the outcomes and safety of assisted reproduction technologies (ART) as a model to address the above, and to explore the capacity to link records of different individuals (for example, mother and child, offspring of the same birth mother) and from different health providers (for example, private fertility clinics and NHS delivered primary or secondary care maternity services).

3) to identify and anticipate developments needed within the Connecting for Health strategy to support such research

1.3 Summary of Methods Used

1) Scoping literature review of studies using record linkage in fields of epidemiology and public health in the UK to highlight existing strengths and opportunities

2) Scoping literature review of studies using record linkage to evaluate outcomes of assisted reproduction technologies to highlight opportunities and strengths

3) Assessment of necessary attributes of relevant electronic health record linkage systems for assessment of outcomes of assisted reproduction based on literature review, discussions with Nordic and other colleagues, staff of the Human Fertilisation and Embryo Authority (HFEA) and the Health Informatics Centre in Dundee

4) Mapping UK data sources for observational epidemiology and for studies of the outcomes of assisted reproduction

5) Characterisation of the Secondary Uses Service [SUS] of Connecting for Health (CfH) and its current potential for observational epidemiology, based on reading of CfH documents and web

pages, and meetings with staff of CfH SUS.

6) Synthesis and identification of generic and specific issues and possible solutions through wider discussion and cross referencing with other simulation leads/projects.

2. OPPORTUNITIES AND BENEFITS OF RECORD LINKAGE

2.1 The Role of Observational Epidemiology

Observational epidemiology embraces a range of non experimental study designs, including cross sectional; retrospective and prospective cohorts;and case control. These study designs have a distinguished tradition of contributing to the prevention and control of disease in populations - often before the exact biological mechanism has been identified - through analysis of the temporal, personal, social and geographical correlates of disease and their putative risk factors. Recent advances in population genetics and biomarker discovery have created exciting opportunities to augment information obtained within large-scale population studies through measurement of genotypic, clinical, environmental and life style factors within large samples from diverse populations. Linkage of these data to routine health service, environmental, fiscal and educational data as well as to vital events (births, deaths etc) can further enhance our understanding and strengthen the scientific basis of strategies to maintain the health and wellbeing of the population and prevent and control disease.

The construction of retrospective cohorts and case control studies based on routine data allows complex causal mechanisms to be investigated in a timely fashion without the need to wait to accumulate person years of observation or events, as is the case in prospective studies. Thus these study designs are particularly helpful when dealing with uncommon outcomes or uncommon exposures where very large scale evidence is needed, or for events with a long interval (even a generation) between exposure and outcome.

Observational studies can also be useful to evaluate the effectiveness and safety of new technologies – including operative procedures, devices and diagnostic tests. While some may be amenable to evaluation through randomised trials, it is unlikely that this will be affordable or appropriate in all circumstances. Decisions about the adoption of new fast-changing technologies at either individual or policy level need to be informed by the best available evidence and require data to be assimilated as quickly as possible.[ref1] In a report commissioned

by the Health Technology Assessment programme, Raftery and colleagues reviewed the potential contribution of record linkage and databases to health technology assessment.[ref2] Databases identifying both health technologies and health states were considered the most useful and included those concerned with adverse event reporting, confidential enquiries, disease-only registers and health surveys. A range of issues to be addressed in order to secure their wider use for technology assessment were flagged, including the need to clarify responsibility for the strategic development of databases, improved resourcing, and issues around coding, confidentiality, ownership and access, maintenance of clinical support, and optimal use of information technology.

Similarly retrospective observational epidemiological studies can inform the response to unanticipated public health concerns about the safety of health service interventions (including, for example, medicines, immunisations, operative procedures, medical devices and transfusions) or environmental exposures (including, for example, radon, mobile phones, landfill sites) in a timely fashion.

2.2 Benefits of Record Linkage

There is increasing recognition that linking of records brings diverse and wide-ranging benefits for patient and public health, partly because it provides the ability to carry out timely, large-scale and definitive studies. Some of these benefits are summarised in the box and rehearsed in greater detail below.

BENEFITS OF USING ELECTRONIC RECORD LINKAGE

Improved patient management, patient safety and health of the public
- Improves patient management through data feedback
- Facilitates long term follow up of participants in clinical trials
- Enhances public health surveillance
- Ensures timely responses to public health concerns
- Enables follow up of patients in sensitive settings
- Patient safety
- Identification of environmental hazards

Improved methodological rigour
- Validating data completeness
- Validating data quality
- Avoiding bias and misleading conclusions
- Identifying biases in consent and participation
- Avoiding ecological fallacies

Efficient large scale research infrastructures
- Intergenerational studies
- Twin studies
- Infrastructure for large scale evidence

2.2.1 Improved Patient Management, Patient Safety and Health of the Public

Improved patient management through feedback of data

Mitchell et al carried out a randomised trial of different forms of feedback linked to electronic data capture in primary care and found that feedback with clear patient prompts tended to lead to better management of hypertension.[ref3]

Long term follow up of participants in clinical trials

Routine records have been used to obtain prolonged follow up of participants in randomised trials to minimise attrition bias. For example, in the 27-year follow up of the offspring of mothers who participated in a randomised trial including diethyl stilboestrol, children were not contacted directly. Information about them, masked to original allocation, was obtained from hospitals, general practitioners, and other official sources and all children were traced.[ref4]

In the West of Scotland Coronary Prevention Study [a double blind placebo controlled trial] ,participants were followed for a period of five years post-trial. Electronic record linkage using the Scottish routine health records and linkage system was used to extend that follow-up to 15 years by capturing information on all subsequent hospital admissions and deaths. This enabled validation of follow-up data from both sources over the first five years, and longer term low cost follow-up of trial participants to confirm the safety and ongoing benefit of pravastatin treatment.[ref5]

Enhancing public health surveillance

In England, the anonymous neonatal seroprevalence

surveys of HIV are enhanced through linkage to routine data from birth registration. The additional information obtained on maternal country of birth is used to monitor demographic trends in HIV infection among pregnant women, to monitor the effectiveness of antenatal screening and treatment for HIV and to predict and plan for the impact of the worldwide epidemic on HIV prevalence and incidence in England.[ref6]

Timely response to public health concerns
The analysis of routine data can provide a timely and authoritative response to public concerns about drug or vaccine safety, as for example, when the uptake of combined measles, mumps and rubella vaccine in two year olds declined from 92% in 1995 to 79% in 2003 in the wake of research, purporting a link with autism and bowel disease. Using the unique Danish identification number given to every live-born infant and new resident in Denmark, Madsen et al were able to link MMR-vaccination status obtained from the Danish National Board of Health, to information on the children's autism status from the Danish Psychiatric Central Register, as well as to the Danish Medical Birth Registry, the National Hospital Registry, and Statistics Denmark which were sources of relevant variables and potential confounding factors. Using these data, 537,303 children were identified (representing 2,129,864 person-years), of whom 440,655 (82%) had received the MMR vaccine. Among these children, 316 had received a diagnosis of autistic disorder and 422 a diagnosis of other autistic-spectrum disorders. The authors found no association between the age at the time of vaccination, the time since vaccination, or the date of vaccination and the development of autistic disorder, providing strong evidence against the hypothesis that MMR vaccination causes autism.

Follow up of patients in sensitive settings
Electronic record linkage allows individuals to be followed without direct contact in situations where such follow-up is precluded due to potential disclosure of information: for example, in following up the health of children born to HIV positive women this might disclose maternal HIV status to the child or wider family [ref7] or in following the health of children born following assisted reproduction this might disclose the biological parentage to the child or wider family.

Patient Safety
Analysis of linked routine data provides an important mechanism for monitoring and evaluating drug and vaccine safety and this issue will be addressed in greater detail in the surveillance simulation report. Highlighted here is the potential to enable efficient monitoring of long term outcomes and concordance of disease between blood donors and recipients through a comprehensive haemovigilance system. The Scandinavian Donations and Transfusions (SCANDAT) database combines data on virtually all blood donors and recipients who have been registered at least once in any of the computerized local blood bank databases in Sweden and Denmark since the start of computerised registration in 1966.[ref8] It links individuals and their entire computerised donation and/or transfusion histories and all donor-component-recipient connections to nationwide population and health registers to attain essentially complete follow-up for up to 36 years regarding reproduction, hospital morbidity, cancer, and death. This powerful high quality database contains information relating to more than one million blood donors who contributed more than 15 million donations and 1.3 million recipients who received more than 11 million transfusions. Within the UK, the Transfusion Medicine Epidemiological Review by the National CJD Surveillance Unit and the UK Blood Services identifies cases and traces donors and recipients of any vCJD cases aged >17 years and the fate of all donations is identified by look back.[ref9] While this system allows examination of specific outcomes, it is arguably less powerful than the Scandinavian system, which allows retrospective cohort as well as case control studies.

Identification of environmental hazards
Record linkage provides a powerful means of evaluating environmental hazards and their implications for human health. Examples in the UK include studies of landfill sites and their relation to birth outcomes[ref10], congenital anomalies [ref11] and cancers [ref12], and the relation of trihalomethanes in the water supply and stillbirth.[ref13] These studies require area level identifiers at earlier points in time as well as data across a range of diverse outcomes including obstetric and perinatal outcomes, birth defects and cancers. In the US record linkage systems have been developed linking environmental exposures to childhood cancers [ref14] and environmental hazards geographically to the occurrence of birth defects.[ref15]

2.2.2. Improved Methodological Rigour

Validating data completeness
Reliable data on congenital anomalies is essential for a range of purposes, including assessment of the safety of assisted reproductive technologies, medicines prescribed or taken in pregnancy and environmental hazards, as well as the effectiveness of antenatal screening and newborn programmes. Current systems in the UK depend on passive notifications to the national scheme or to regional registers. In Nordic countries a number of studies have been carried out to evaluate the completeness of data on congenital anomalies using different methods. For example, Larsen compared data in the Danish Medical Birth, Hospital Discharge and National Congenital Abnormalities registries and reported that Hospital Discharge data were most sensitive and specific in identifying congenital anomalies.[ref16] In a similar study in Finland, Hemminki also concluded that routine hospital data were the most complete and specific across all groups of malformations and were more useful for routine surveillance of malformations.[ref17] In both studies, the coverage of malformation registers was found to be very low.

Validating data quality
There are numerous examples of the use of data linkage to validate the quality of specific datasets where neither can be considered the 'gold standard'. For example, children participating in the UK Millennium Cohort study were linked to their birth registration records to compare maternal report of offspring birth weight with that recorded in the birth registration record.[ref18] Agreement to within 100 g was 92% overall, but this was lower in minority ethnic groups, the long-term unemployed and those living in disadvantaged or ethnic areas, with over one quarter of the discrepancies of 100 g or more identified as due to errors of recording or transcription. The need for cross validation of routine data sources has been highlighted by recent studies examining mortality after surgery for congenital heart defects, using Hospital Episode System data [ref19] or data from the Central Cardiac Audit Database.[ref20] These studies reached some differing conclusions regarding mortality risk in specific centres, highlighting the importance of validating completeness and accuracy of data sources and of being able to examine patient identifiable data to do so.

Avoiding bias and misleading conclusions
Misleading conclusions may arise where there is differential misclassification of exposures. For example, a collaborative reanalysis of data from 53 epidemiological studies, including 83,000 women with breast cancer, was conducted to determine whether induced abortion led to breast cancer. The relative risk of breast cancer was found to be increased for women having had a pregnancy that ended as an induced abortion when this exposure was recorded retrospectively, but decreased when based on studies recording this prospectively (including several using routine linkage). This was possibly because women who had developed breast cancer were on average more likely than other women to disclose previous induced abortions.[ref21] Record linkage to prospectively recorded information, where this is possible, thus it can avoid the problem of recall bias in epidemiological studies.

Identifying biases in consent and participation
Proposals to seek individual consent to record linkage of routine data have been advanced. In an analysis of parents who did and did not consent to such linkage of their child's health records in the Millennium Cohort Study, Tate et al demonstrated non-consent bias, with lone parents, older mothers, and those from certain minority ethnic groups less likely to consent.[ref22] Anonymised record linkage of all eligible subjects is thus less likely to exclude certain sectors of the population and can avoid biases due to differential social exclusion.

Avoiding ecological fallacies
Donnan et al analysed resistance to trimethoprim (an antibiotic) in bacteria obtained from urine samples at practice and individual level simultaneously in a multilevel model. Prescribing and antibiotic resistance were not associated at practice level, but were significantly associated at an individual level, highlighting the added value of individual patient data for research on the outcomes of prescribing and the potential for reaching misleading conclusions in ecologically based analyses.[ref23]

2.2.3. Efficient large scale research infrastructure

Intergenerational cohorts
Record linkage can be used to construct intergenerational cohorts. For example, Hypponen et al linked the parents of 1958 birth cohort members to their mortality data in order to evaluate multigenerational associations with health

and development in parents, their offspring and grandchildren.[ref24] Record linkage has been used in the General Practice Research Database to link mothers to their offspring [ref25,26], and siblings and to identify siblings and multiple births in an examination of the risk of allergic disease.[ref27]

Twin registers
Twin registers are a powerful resource for genetic epidemiology. Record linkage has been exploited to create twin registers, most notably in the Danish twin register which is one of the oldest and most established spanning 127 birth cohorts.[ref28] Similar systems have been employed or explored in Western Australia, Sweden and Italy.[ref29-31]

Infrastructure for large scale evidence
Very large samples are needed to address the predictors of uncommon outcomes or to examine specific associations with uncommon exposures. For example, Scottish record linkage systems have been successfully used to assess or predict the risk of adverse pregnancy and perinatal outcomes in successive pregnancies, including the risk of sudden infant death and unexplained stillbirth.[ref32;33] Lawlor et al [ref34] used linked data from 1.8 m Swedes to examine the contribution of childhood socioeconomic position to adult cause specific mortality for which previous studies had lacked sufficient power. She found that, with the exception of stomach cancer caused by Helicobacter pylori infection acquired in childhood, poorer social class in early life was associated with diseases largely caused by modifiable behavioral risk factors such as smoking, physical inactivity, and an unhealthy diet.

2.3 Assisted Reproductive Technologies

2.3.1 Rational

The above brief review highlights the scope of observational epidemiology based on routine data linkage. In conducting the simulation exercise we elected to focus on the use of record linkage to evaluate outcomes of assisted reproductive technologies. We considered that this model would draw out a number of elements of generic relevance to retrospective cohort and case control studies and would serve as a model for epidemiological research, with many elements generalisable to questions of disease cause, longer term outcomes ('natural history') and non-experimental study designs for technology assessment. Some of these generic elements are described in more detail below.

This model is relevant to research addressing the *longer term implications of interventions and exposures in pregnancy* and early life for maternal, child and intergenerational ill health, health and development, as well as for questions concerning the early life origins of adult disease. Assisting reproduction involves *fast-changing technologies* which are rarely evaluated in controlled trials.

In addition, ART as a model has timely relevance to a range of current national and international research priorities and agendas in maternal and child health.

The *effectiveness and safety of all new assisted reproductive technologies*, including their short and longer term health outcomes for mother and child, was highlighted as a key priority by the joint MRC and HFEA review[1] of assisted reproductive technologies and endorsed subsequently by the Parliamentary Science and Technology Committee. At present evaluation of ART is complicated by current legislation, which proscribes record linkage. Thus there has been virtually no useful large scale epidemiological research addressing the outcomes of ART in the UK, despite more than 1.5 million cycles of treatment recorded in the HFEA database which was started in 1991.[ref35] Legislative review of the Human Fertilisation and Embryology Act is now in an advanced stage with a declared intention of removing legal bars to future research of offspring and maternal health following ART. It is therefore timely to ensure that there is the capacity to realise that intent and to support such research through the National Care Record Service as well as through the HFEA (shortly to be merged with the Human Tissue Authority), a significant data custodian as the regulator of the fertility clinic sector. Furthermore, the research potential of the existing database going back to 1991 has never been realised and the possibilities of linkage – legal, technical and operational - should be explicitly addressed given the potential person years of follow up represented.

Many of the issues raised through ART are relevant to addressing the *longer term health, educational and social outcomes of prematurity* as highlighted in the recent Nuffield Council report.[2]

1) Assisted reproduction: a safe sound future. Medical Research Council, 2004
2) Critical care decisions in fetal and neonatal medicine. Nuffield Council on Bioethics, 2006

Similarly the model is relevant to the *evaluation of existing and proposed antenatal and newborn screening programmes*, a number of which have been the subject of NHS CRD Health Technology Assessment Programme priorities. Related to this is the evaluation of the health and wellbeing of children born following preimplantation genetic diagnosis and the potential for health gain of predictive genetic testing of the newborn which were the subject of two recent Human Genetics Commission reports.[3,4]

The *implications of pharmaceutical treatments during pregnancy for offspring health* in the short and longer term are an important consideration for drug safety. This includes conventional prescribed and alternative medicines.[ref25;36;37] Additional examples currently under review include the longer term effects on uninfected offspring of highly active antiretroviral treatment in HIV positive pregnant women.

Time trends in, and geographical clustering of, congenital malformations and other rare disorders and their relation to individual and environmental factors, have received recent attention in the Chief Medical Officer's Report,[5] where it was noted that better surveillance systems were needed.

Finally the importance of identifying the *reproductive implications of chronic conditions*, for example, diabetes and obesity, was emphasised in a recent report of the Confidential Enquiry into Maternal Deaths.[6]

As a model ART allows some specific issues particularly pertinent to children to be explored which may not be covered within other simulations, including:

- linkages between individuals (e.g. mother to child)
- linkage with data from private sector providers
- linkage with other data sources including those relevant to education, social care, and environmental hazards
- the extent to which these are possible for both exposed and unexposed (control) populations.

2.3.2 Scoping Literature Review of Record Linkage for Art Outcomes

It is estimated from the WHO report on assisted reproductive technologies that about one million children have been born worldwide after IVF treatments. The number born after other forms of ART is estimated to be similar. In the last decade data on the percentage of births in Europe following ART has been monitored by the European Society for Human Reproduction and Embryology [ref38]. In 2001, the proportion of infants born after ART with in vitro techniques ranged from 0.2% in Latvia to 3.9% in Denmark. In general, the impact of the in vitro techniques on the birth rate was reported to be highest in the Nordic countries (range 2.2-3.9%) reflecting their wider availability. Denmark has an established national ART reporting system involving all ART treatments: in 2002, 6.2% of all infants were born after assisted reproduction, 2% following intrauterine insemination and the remainder following in vitro techniques. Comparison of data between countries has been facilitated by the development of a common international dataset of definitions and terminology by the International Committee Monitoring Assisted Reproductive Technologies.[ref39] This facilitates the dissemination of ART data and its availability, efficacy, and safety to health professionals, health authorities, and the public.

Explicit systems have been established, or opportunistic studies carried out, based on record linkage to evaluate the later maternal and infant health outcomes of ART in a number of countries, including Sweden[ref40], Denmark[ref38], Finland[ref41], and Canada[ref42]. Other countries have employed alternative methods, including aggregate annual reporting by fertility clinics of short term outcomes against an agreed data template to a central system as in the United States[ref43] or Sweden[ref40], specific multicentre studies with individual clinical follow up such as in Belgium [ref44], or collection of individual level data from clinics such as in the UK. [ref35]

Sweden has established two permanent and independent systems to monitor ART outcomes.[ref40] These will be described in detail as they encompass

3) Profiling the newborn: a prospective gene technology? Human Genetics Commission 2005
4) Making babies: reproductive decisions and genetic technologies Human Genetics Commission 2006
5) On the State of the Public Health: Annual Report of the Chief Medical Officer 2004, DoH, 2005
6) Why mothers die 2000-2002. 6th report of the Confidential Enquiries into Maternal Deaths in the UK, CEMACH, 2004

the two main models in use internationally: one based on annual aggregate reports and the other based on surveillance and record linkage. Since 1987, Swedish law has required all fertility clinics to submit annual summary reports on the results of treatments starting during a specific year and including the outcome of the resulting pregnancies. These reports are sent to the Centre for Epidemiology at the National Board of Health and Welfare which is responsible for the preparation of an annual national report. The performance of individual clinics is not reported, in contrast to the UK, out of concerns that this may be misinterpreted by the public.

A separate surveillance system has been established to collect health outcome data for children born since ART procedures were introduced in Sweden.[ref45] This system operates through separate reporting from all clinics on all women who have delivered one or more children after ART treatment. Each Swedish citizen has a unique ten-digit identification number (PIN) and the clinic report includes each woman's identification number. The Centre for Epidemiology has developed a specific register covering all ART deliveries in Sweden and this ART delivery register is linked to the Swedish Medical Birth Registry, the Cancer Registry and the Registry for Malformations. These also use the same PINs and, therefore, each woman can be identified in all registers. All deliveries in Sweden are reported to the Swedish Medical Birth Registry, which has been in existence since 1973 and which collects information for the pregnancy, delivery and the immediate postpartum period based on standard medical documents used in all Swedish antenatal care centres, delivery and neonatal units. The PIN for the ART children is identified through this Medical Birth Registry which therefore links mothers to their children.

In Denmark, all ART activity using in vitro techniques has been reported on a statutory basis to the Danish National Board of Health since 1994, and since 1996 also to The Danish Fertility Society.[ref38] Treatments with intrauterine inseminations have been reported on a voluntary basis to the Danish Fertility Society since 2001. Individual treatment cycles are reported to the National Board of Health and identified using the personal identification number (CPR number) which can then be used to link to other routine health registers in Denmark. This unique identifier is a major attribute of both the Swedish and Danish systems.

Both Sweden and Denmark have published extensively on the child and maternal outcomes of ART using these surveillance and reporting systems combined with record linkage. Gissler has highlighted the enormous potential of the Nordic birth registries,[ref46] and linkage to malformation, cancer, hospital and psychiatric registries allows evaluation of a wide range of outcomes, enables multiple control groups to be identified, and adjustment for potential confounding factors to be made in the analyses. For example, Bergh[ref45] was able to investigate, in a retrospective study, malformations, cancers, and deaths in the complete Swedish in-vitro-fertilisation (IVF) birth cohort compared with the general population using data from the Swedish Medical Birth Registry, the Registry of Congenital Malformations and the Swedish Cancer Registry with stratification for maternal age, parity, previous subfertility, year of birth, and multiple of pregnancies. Most of the adverse outcomes observed were attributed to multiple births and maternal characteristics. Only 0.9% of IVF births could not be identified as delivered. This study highlights the importance of individual level data on potential confounding factors as well as the completeness of linkage achieved.

Subsequent studies have provided ongoing evaluations of the risks of congenital malformations, childhood morbidity, and childhood cancer as well as maternal mortality and morbidity using the same linked datasets, demonstrating the effectiveness and dynamic nature of the system for research.[ref47-50] In a study of maternal morbidity a potential increased risk of ovarian cancer was reported: one strength of this study was the ability to exclude cancers with onset that predated IVF treatment.[ref47] A further important attribute of the Swedish data is the ability to link to laboratory data: Kallen examined the risk of congenital malformations following different IVF methods, combining analyses of child and maternal outcomes, laboratory data and potential parental confounding factors.[ref50,48]

A large number of studies of ART outcomes have been reported from Denmark, including studies examining maternal risks and perinatal outcomes[ref51], neonatal outcomes following different ART methods[ref52;53], the frequency of rare disorders of genetic imprinting[ref54;55], cerebral palsy[ref56], neurological sequelae in twins[ref57], and hospital care utilisation by offspring conceived following ART[ref58].

While most of these studies have reported relatively short term outcomes, more recent studies have been able to address longer term maternal outcomes taking into account potential confounding by indication. Mosgaard[ref59] assessed the risk of invasive ovarian cancer among infertile women treated with fertility drugs using registers to identify all Danish women below the age of 60 years with ovarian cancer during a defined period and age-matched population controls. After allowing for the influence of important confounding factors such as parity and infertility, treatment with fertility drugs did not increase the ovarian cancer risk compared with nontreated infertile women among parous as well as nulliparous women. This study highlights the importance of being able to select, using routine health registers, an appropriate control group to allow for potential confounding.

The Danish record linkage system also allows siblings to be linked, although this has not, as far as can be established, been employed to date in ART. In a study of sibling risk following diagnosis of childhood cancer, Winther[ref60] identified children with cancer from Nordic cancer registries, and their siblings from nationwide population registries and documented cancers in siblings through record linkage with cancer registries and compared cancer rates with national incidence rates. Cancer incidence in parents was also examined to identify familial cancer syndromes. The authors concluded that, apart from rare cancer syndromes, paediatric cancer is not an indicator of increased cancer risk in siblings.

Research based on Nordic registers provides an opportunity to assess the quality of the data recorded and in general this has been high within the specific studies reviewed for this exercise. Gissler has reviewed a number of Finnish registers and identified the need to dynamically maintain and evaluate data quality using a variety of mechanisms.[ref61-63] Although some registers, such as the Danish register of psychiatric diagnoses, have been politically controversial at times[ref64], rigorous standards to safeguard confidentiality and anonymity have ensured that their value is commensurately greater than any risks of disclosure.

2.3.3 Summary of attributes of record linkage systems to address outcomes of ART

The above brief review has highlighted several important components of record linkage systems used for observational epidemiological studies of the outcome of ART. These include:

- Complete and inclusive coverage of all assisted reproductive treatments
- Use of a unique identifier to allow individuals to be linked across different registries and data sources
- Ability to link unique identifiers by families, namely mother to child, father to child, and sibling to sibling
- Availability of authoritative complete and high quality sources of a wide range of outcome data, including perinatal events, congenital malformations, neurological outcomes such as cerebral palsy, cancers, psychiatric outcomes, mortality and other morbidity
- Availability of relevant confounding or other factors, namely, multiple births, gestation, parity, infertility history
- Availability of laboratory and other data to allow evaluation of outcomes in relation to specific technologies
- Ability to select a varied range of control groups to test consistency of findings or to examine specific hypotheses
- The value of historic or 'legacy' data in reducing the time needed to accumulate person years of observation and adequate power to test specific hypotheses.

In the next section we explore the extent to which these attributes are available within current UK systems and within the proposed Secondary Uses Service.

3. CURRENT SITUATION

3.1 Current Data Sources for Observational Epidemiology in England

Observational epidemiology in England currently relies on a wide range of opportunistically linked datasets and resources. These have been 'mapped' and are shown in Figure 1 together with their potential relation to data within the NHS Care Record

Service.

The Office of National Statistics has provided and continues to provide high quality data from registration of births, marriages and deaths. Certain data, notably that on termination of pregnancies, is safeguarded to avoid deductive disclosure and is only released in aggregate form, as for example, for terminations for congenital rubella which are uncommon (Personal communication, Dr Pat Tookey). The links between vital statistics data and cancer registers are relatively well developed and the 'flagging' service at Stockport allows individuals in a range of epidemiological surveys to be traced and flagged in relation to later migration or death, given the correct permissions.

Hospital episode data in England have been transformed from episodes to individuals, but completeness and data quality remain an issue as exemplified in the recent record linkage exercise to the maternity Hospital Episode System (HES) data tail carried out for the Millennium Cohort Study. Overall a matching hospital record was found for 83% of those who consented but this varied markedly across the UK (82% in England, 91% in Wales, 93% in Scotland and 66% in Northern Ireland). However completeness of data showed marked variation, notably in relation to gestational age: for example, in England gestational length was available for 83% of singleton infants whose parents consented to linkage and for whom a matching record was found: 94% of these records agreed with maternal report of gestation to within 2 weeks either side. These observations on quality of gestational age data are consistent with recent reports on the data completeness of the 'maternity tail' of HES data.

Overall data in Scotland, linked using probabilistic rather than deterministic methods, was of higher completeness and quality. The Scottish system of databases and record linkage has much in common with the Nordic systems, originating in the mid 1970's with the Community Health Index or 'CHI' number, a unique identifier which has been increasingly used throughout all health systems in Scotland. Although the completeness of its use varies across Scotland, it is sufficiently high that, combined with the use of probabilistic linkage methods, the coverage and completeness of data obtained through routine linkage is high. As part of this simulation exercise, a number of members of the simulation teams visited the Information Centre on Tayside where Professor Andrew Morris and colleagues outlined the elements of this system, including the use of record linkage to construct a large intergenerational cohort in Dundee – the Walker project,[ref65] as well as for Generation Scotland. It is clear that Scotland has developed an enviable system and has much experience and expertise to share with other UK countries.

Primary care databases have enormous research potential which has been realised through a number of systems including QResearch which is based on Egton Medical Information Systems (EMIS) a computer system used by the majority of UK general practitioners[ref66], the General Practice Research Database (GPRD) which is owned by the Medicines and Healthcare Regulatory Authority and which covers about 5% of general practices in England[ref67], and others including the Doctors Independent Network[ref68;69]. QResearch has been used to examine the relation of deprivation and ethnicity to quality of care[ref70], to examine the safety of cyclo-oxygenase 2 inhibitors[ref71;72], to examine the effectiveness of secondary prevention of mortality from ischaemic heart disease[ref73] and to assess the contribution of hormone replacement therapy to ischaemic heart disease in women[ref74]. In a recent development, a system to assist in monitoring the risk of pandemic flu – Qflu – has been developed which provides a mode of active public health surveillance.[ref75] The GPRD as been extensively used to evaluate drug safety, and has, in addition, been used to identify pregnancies[ref26], to link mothers to their offspring([ref25], to identify individuals with congenital malformations such as neural tube defects[ref76] and to link siblings.[ref27] The Doctors Independent Network has been compared with the GPRD in relation to a range of common childhood conditions: both systems were found to work better for well defined clinical conditions, with some advantages being reported for childhood eczema for the former system which is based on a problem oriented medical record approach[ref68;69].

3.2 Current UK Data Sources for Epidemiological Studies of Art

Figure 1 was reviewed and data sources relevant to the evaluation of ART highlighted in red [Figure 2]. It is clear from both these figures that the data sources for observational epidemiology, whether or not explicitly for ART evaluation, are extensive and

Figure 1 Data Sources for Observational Epidemiology

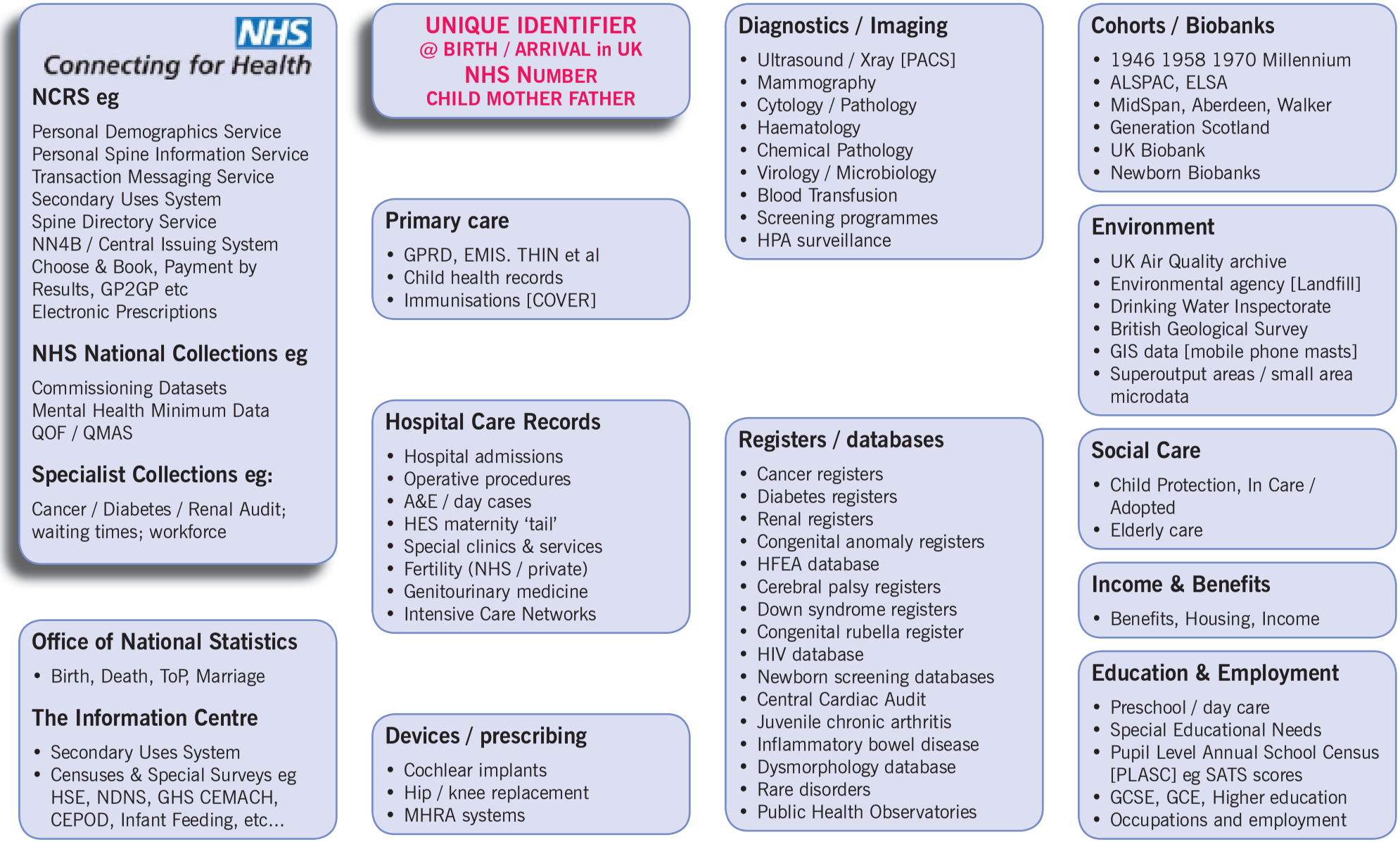
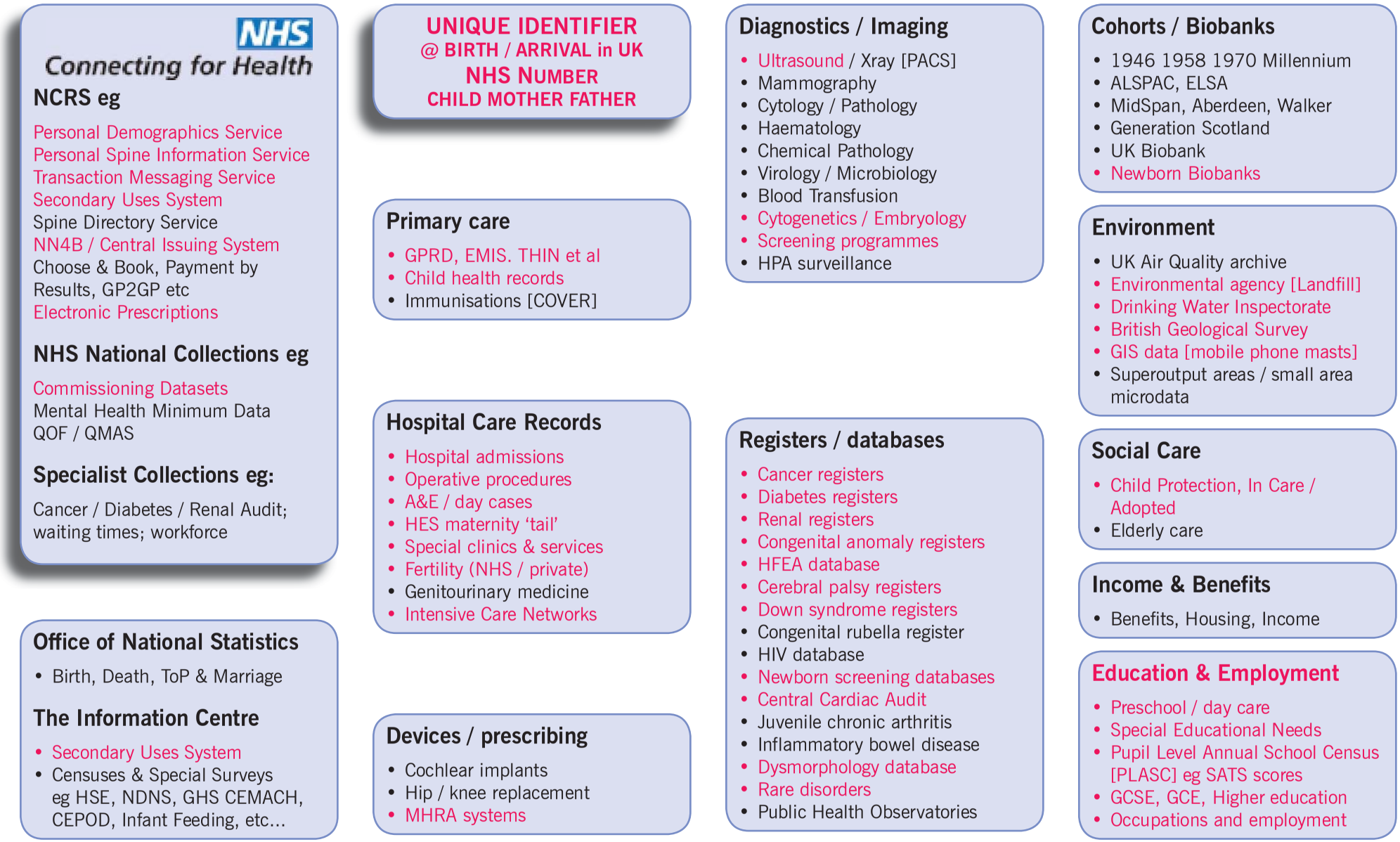


Figure 2 Data Sources for Infertility and ART Epidemiology



that a significant number are outside the traditional boundaries of health care and the NHS. It is also clear that a number of the essential data sources are lacking a sustainable resourced infrastructure or are of variable quality

A key concern for ART is the link with data held in the private sector. Currently, in general, the private sector is only required to report data for services commissioned by the NHS and for regulatory purposes. Fertility services are provided by the private sector to around 70% of couples and these clinics contribute to the HFEA database. The existing database was established in 1991 and has been extensively reviewed and internally validated. By law this database cannot be linked to other data except by the HFEA itself and therefore little is known regarding its external validity. It comprises around 1.5 million records with each record representing a cycle of treatment and, until recently, lacked any person level identifier. Data on offspring are very incomplete and NHS numbers are also incomplete. The HFEA has recently employed probabilistic methods to link records and create a person level identifier but this has not, to our knowledge, been validated externally. This legacy dataset is an important dataset for the field but it is unclear whether the forthcoming legislative review will relax the constraints on linkage.

In 2005 the HFEA introduced electronic data capture systems with the fertility clinics but the data collected are primarily for regulation and less information is now collected on laboratory procedures which might be needed to explain any variations in outcome. Other changes include the allocation of a person level identifier. However mechanisms for ensuring follow up of offspring are not well articulated.

Existing research on ART in the UK is therefore very limited and the information contained in the HFEA database has not been fully exploited. The original MRC Working Party on Assisted Reproduction reported on the adverse consequences of multiple embryo transfer, multiple pregnancies and prematurity[ref77] and this was endorsed in a subsequent analysis of data from the HFEA database[ref78] One small study based on the GPRD examined the association of IVF with retinoblastoma but the low rate of assisted pregnancies identified suggested underascertainment.[ref79] UK scientists have developed small cohorts of children in the context of wider European studies and followed their early child health and development but this has not capitalised on the data held centrally to any extent.[ref80-83] Studies of the longer term outcomes of ovarian stimulation in the UK are lacking and it is notable that this treatment is not regulated by the HFEA. Reproductive epidemiological studies have been undertaken more broadly in the UK, highlighting the capability in selected databases for the study of environmental and other determinants of outcome.[ref 10;12;13;27;32;33;84-86] These have highlighted lack of high quality data on gestation and congenital anomalies, as well as problems of accessing data linked to postcode for environmental epidemiology. (personal communication, Paul Elliott).

3.3 National Care Record Service and Secondary Uses Service

The NHS Care Record Service comprises a spine with different components which are summarised in the box.

What are the different parts of the Spine and what do they do?*

- **Personal Demographics Service**
  - a single central source for patient demographic information
- **Personal Spine Information Service**
  - the central database containing clinical records for each NHS patient
- **Transaction Messaging Service**
  - allows clinical messages from NHS CRS users to be securely routed
- **Secondary Uses Service**
  - provides timely, anonymous patient data and information for purposes other than direct clinical care, while protecting patient confidentiality
- **Clinical Spine Application**
  - provides healthcare professionals with access to the NHS CRS to gain controlled access to and update patient information provided by the PDS and the PSIS.
- **Spine Directory Service**
  - Look up service comprising the Spine User Directory and Spine Accredited Systems and Services

- **Access Control Framework**
  - registers and authenticates all users accessing a patient's NHS Care Record [Role Based Access and 'sealed envelopes']

**adapted from spine factsheet produced by CfH, accessed from website 21.1.2007*

The Secondary Uses Service is the proposed route for access to data for purposes other than direct clinical care. Currently information about SUS provided on the CfH website does not specify research but states 'The Secondary Uses Service (SUS) will protect patient confidentiality and will provide timely, anonymous patient data and other information for purposes other than direct clinical care. This includes looking at public health trends, analysing the effectiveness of treatments and planning the number of beds and staff the NHS needs. SUS will support a number of national initiatives, the first being Payment by Results, a key government initiative which is changing the way money flows through the NHS.' This[7] gives the impression that business uses and aggregated data for business or health service planning is foremost.

Currently the main datasets within the NHS CRS of value to retrospective observational epidemiology include:

- The Central Issuing System / NHS Numbers for Babies which allocates the NHS number at birth or at immigration
- The Commissioning Datasets which have specific records for birth and delivery episodes, which include a minimum dataset for neonatal intensive care
- The Personal Demographic Service
- The Personal Spine Information Service

The relation of these to SUS is shown in Figure 3, and further detail of SUS shown in Figure 4, both reproduced with permission from Jon Fistein and Simon Heathfield, its authors.

SUS business needs are stated as:

Access and Choice
in support of capacity and demand planning, commissioning, linked to the implementation of Payment by Results

Standards and Performance Monitoring
looking at National Service Frameworks, clinical indicators, and information to support the work of the Healthcare Commission

Public Health information
including screening, surveillance and epidemiology. In particular this would support the Public Health Information Strategy, announced in the Choosing Health White Paper (2005)

Research and Development
including longitudinal studies and the monitoring of outcomes and effectiveness supporting a range of development activities

Improving Productivity
in areas such as the GMS contract, the new consultant contract, Agenda for Change and benchmarking.

Further details regarding functionality highlight a concept of "*one national approach to the Secondary Uses Service with user access managed through the security and confidentiality facilities embedded within NHS CRS. Information provided through the Secondary Uses Service will be pseudonymised and, where possible, will be collected or derived from clinical systems as a by-product of direct care; SUS will include the tools and services for an effective and secure working environment for analysis and reporting.*"[8]

The literature review, mapping exercise and simulations we have conducted suggest that there are potential opportunities within SUS to deliver some functionality for research but that of itself SUS is not sufficient to ensure a strong and vibrant research industry based on record linkage.

We explored in some detail the ability to identify and link children to their mothers using the NHS number

7) spine factsheet produced by CfH
8) Simon Heathfield powerpoint presentation to December RCGP 2006

Figure 3 The NHS Care Records Service

Access to data collected in operational care for research, planning...

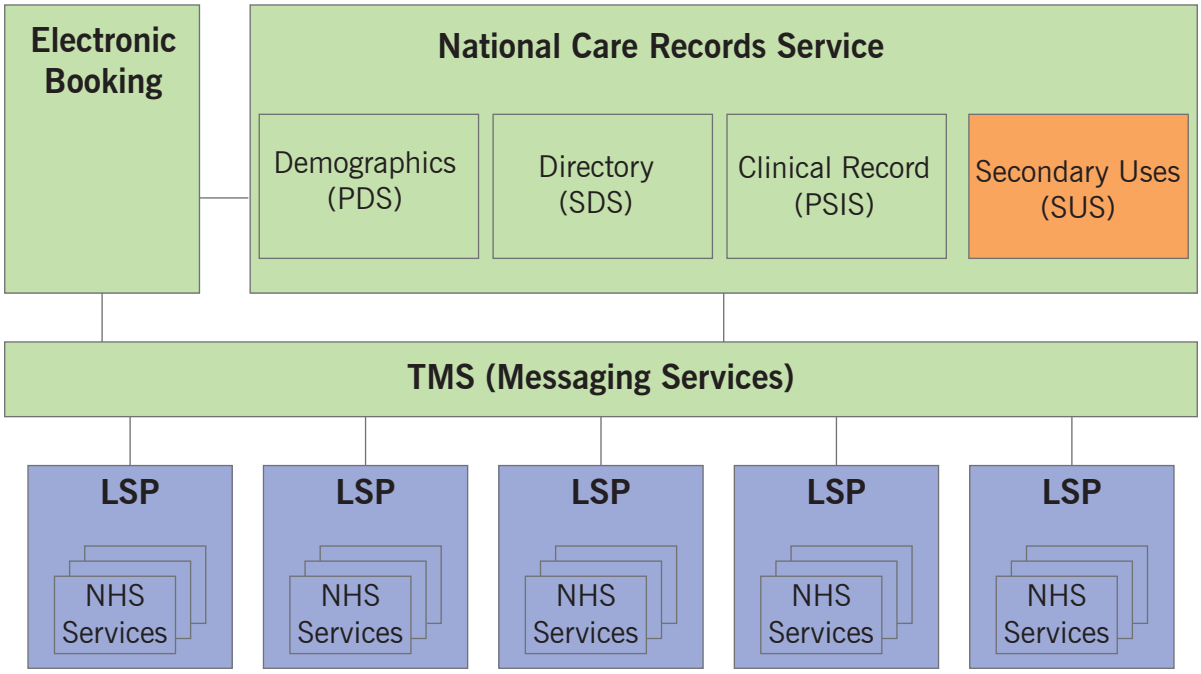
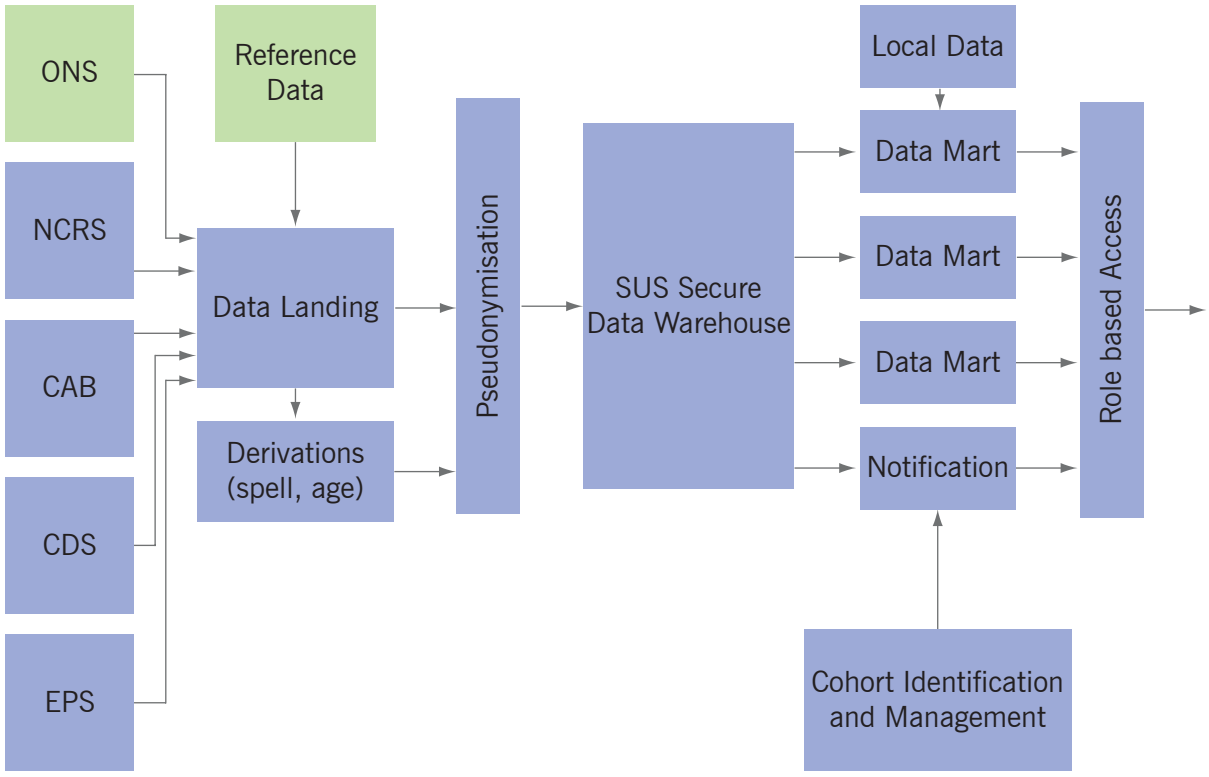


Figure 4 Secondary Uses Service Architecture



within four datasets currently within the SUS – the birth notification dataset, the commissioning datasets for birth and delivery episodes and the personal demographic service (Table). We also identified the completeness of data for other key variables for ART, notably multiple births, birthweight, gestation, parity and congenital malformations. We noted which variables were cross indexed to Hospital Episode System data.

All datasets included NHS number, although this was only populated in the birth notification dataset after notification of the birth by the midwife. However this variable was recorded as optional in the commissioning datasets according to the on-line data dictionary. The potential to link children to their mothers was identified in that the maternal NHS number is recorded in the birth notification dataset as well as the commissioning dataset for the delivery episode and birth episode (optional in the latter). The NHS number is cross referenced to the Hospital Episode system data. Thus the potential is there to have a unique identifier and to link mothers to children (and potentially to all their children).

Congenital malformations were only recorded in the birth notification dataset and the extent and completeness of these is unclear. It is our understanding that the birth notification dataset is only retained for a limited period and this needs to be clarified.

Linkage of these datasets to primary care is not yet possible at an individual level although we understand some of the primary care databases are developing some linkages with hospital episodes, mortality and area level indices of deprivation. SUS has no immediate plans for direct linkage to primary care or pathology services, both of which are essential for assessment of studies of disease outcomes as well as technology assessment.

The uncertain legal status and lack of information about the quality of data in the current HFEA database make it difficult to assess the potential for utilising this legacy dataset.

Table: Comparison of identifying data and essential data items across four datasets tabulated by Connecting for Health Data Dictionary and NN4B

	NN4B birth notification dataset	CfH PDS	CDS: delivery episode	CDS: birth episode	HES cross referenced
NHS No	Once notified	✓	Optional?	Optional?	✓
Name		✓		Optional	
Ethnic category	✓			Optional	
Live/Stillbirth	✓		✓	✓	✓
Date of birth (baby)	✓	✓	✓	✓	✓
Birth order	✓	✓		✓	✓
Number of babies	✓		✓	✓	✓
Birth weight	✓		✓	✓	✓
Gestation	✓		✓	✓	✓
Delivery method			✓	✓	✓
Date of death		✓			
Death notification		✓			
Sex	✓	✓	✓	✓	✓
Address, postcode	✓	✓	✓	NO*	✓
Child health organisation code	✓				
Code of GP practice	✓		Optional	Optional	
Neonatal care level				✓	
1o/2o diagnoses [ICD]			✓	✓	✓
1o/2o diagnoses [Read]			Optional	Optional	
Procedure codes [OPCS]			✓	✓	✓
Suspected congenital anomaly	✓				
NHS Number (mother)	✓		✓	Optional	✓
Birth date (mother)	✓		✓	Optional	✓
Mother ethnic category			✓		✓
Postcode			✓	✓	✓
Total previous pregnancies				✓	✓
Total previous live births (parity)	No	No	No	No	-

4. PROPOSED SOLUTIONS

4.1 In the light of this review and simulation exercise, we consider that an electronic record research infrastructure fit for observational epidemiology requires a *high level shared and integrated vision* of data for clinical management, performance management and audit, epidemiological research and clinical trials. This requires an explicit resourced and governanced policy for linking multiple datasets

4.2 The main role of the Secondary Uses Services should be to underpin a *process* linking the strategic elements and assuring sustainability, quality and governance of the linkage process and not as a data warehouse.

4.3 A proactive strategy for linking primary care, laboratory services and modern disease registers to birth registers and hospital data is not in place and we suggest that this is addressed as a matter of priority.

4.4 The UK has many strengths in using record linkage for epidemiological studies and these need to be built on. They could be significantly enhanced by mandating the use of the unique identifier – the NHS Number- in all key NHS records and activities. Changing the recording of maternal and child NHS number from optional to mandatory in key SUS datasets would enhance ability to link records prospectively and allow future mothers to be linked to their children. This strategy needs to embrace deliveries in the private sector.

4.5 A different strategy will be needed to enhance the use of the NHS number in hospital and other datasets and to improve the completeness of linkage in legacy datasets. At present we were not able to establish the extent to which historic data would be incorporated into the SUS.

4.6 A strategy for ensuring data quality and completeness is required and the contribution of research to its evaluation and improvement should be acknowledged.

4.7 The role of multiple source active disease surveillance using electronic records needs to be developed, for example, in relation to congenital anomalies and the role of disease registers needs to be revisited.

4.8 Data on private sector treatment and outcomes should be integrated within the NHS system and mechanisms to achieve compliance with this requirement explored.

4.9 Children are not small adults: a shared information strategy with Children's Trusts is required to capture community, environment, education and social care data.

4.10 The quality and completeness of specific data items essential for ART evaluation as well as other types of study needs to be addressed, namely gestational age and congenital anomalies.

4.11 Routine collection of data on parity should be introduced.

4.12 Mechanisms to increase the completeness and quality of data returns for the maternity 'tail' of the Hospital Episode System should be identified.

4.13 The quality of the HFEA legacy database as a linked research resource should be examined and, if appropriate, representations made to legalise its use for this purpose

Reference List

1. Bull C, Yates R, Sarkar D, Deanfield J, de Leval M. Scientific, ethical, and logistical considerations in introducing a new operation: a retrospective cohort study from paediatric cardiac surgery. BMJ. 2000;320:1168-73.
2. Raftery J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. Health Technol.Assess. 2005;9:1-106.
3. Mitchell E, Sullivan F, Grimshaw JM, Donnan PT, Watt G. Improving management of hypertension in general practice: a randomised controlled trial of feedback derived from electronic patient data. Br.J Gen.Pract. 2005;55:94-101.
4. Beral V,.Colwell L. Randomised trial of high doses of stilboestrol and ethisterone therapy in pregnancy: long-term follow-up of the children. J.Epidemiol.Community Health. 1981;35:155-60.
5. Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. The West of Scotland Coronary Prevention Study Group. J.Clin Epidemiol. 1995;48:1441-52.
6. Ades AE, Walker J, Botting B, Parker S, Cubitt D, Jones R. Effect of the worldwide epidemic on HIV prevalence in the United Kingdom: record linkage in anonymous neonatal seroprevalence surveys. AIDS. 1999;13:2437-43.
7. Hankin, C., Lyall, H., Peckham, C. S., and Tookey, P. Monitoring death and cancer in children born to HIV-infected women in England and Wales: use of HIV surveillance and national routine data. AIDS . 2007. In Press
8. Edgren G, Hjalgrim H, Tran TN, Rostgaard K, Shanwell A, Titlestad K et al. A population-based binational register for monitoring long-term outcome and possible disease concordance among blood donors and recipients. Vox Sang. 2006;91:316-23.
9. Hewitt PE, Llewelyn CA, Mackenzie J, Will RG. Creutzfeldt-Jakob disease and blood transfusion: results of the UK Transfusion Medicine Epidemiological Review study. Vox Sang. 2006;91:221-30.
10. Morris SE, Thomson AO, Jarup L, de Hoogh C, Briggs DJ, Elliott P. No excess risk of adverse birth outcomes in populations living near special waste landfill sites in Scotland. Scott.Med.J. 2003;48:105-7.
11. Dolk H, Vrijheid M, Armstrong B, Abramsky L, Bianchi F, Garne E et al. Risk of congenital anomalies near hazardous-waste landfill sites in Europe: the EUROHAZCON study. Lancet. 1998;352:423-7.
12. Jarup L, Briggs D, de Hoogh C, Morris S, Hurt C, Lewin A et al. Cancer risks in populations living near landfill sites in Great Britain. Br.J Cancer. 2002;86:1732-6.
13. Toledano MB, Nieuwenhuijsen MJ, Best N, Whitaker H, Hambly P, de Hoogh C et al. Relation of trihalomethane concentrations in public water supplies to stillbirth and birth weight in three water regions in England. Environ.Health Perspect. 2005;113:225-32.
14. Hanrahan LP, Anderson HA, Busby B, Bekkedal M, Sieger T, Stephenson L et al. Wisconsin's environmental public health tracking network: information systems design for childhood cancer surveillance. Environ.Health Perspect. 2004;112:1434-9.
15. Brender JD, Zhan FB, Suarez L, Langlois P, Gilani Z, Delima I et al. Linking environmental hazards and birth defects data. Int.J Occup. Environ.Health. 2006;12:126-33.
16. Larsen H, Nielsen GL, Bendsen J, Flint C, Olsen J, Sorensen HT. Predictive value and completeness of the registration of congenital abnormalities in three Danish population-based registries. Scand.J.Public Health. 2003;31:12-6.
17. Hemminki E, Merilainen J, Teperi J. Reporting of malformations in routine health registers. Teratology. 1993;48:227-31.
18. Tate AR, Dezateux C, Cole TJ, Davidson L. Factors affecting a mother's recall of her baby's birth weight. Int.J Epidemiol. 2005;34:688-95.

19. Aylin P, Bottle A, Jarman B, Elliott P. Paediatric cardiac surgical mortality in England after Bristol: descriptive analysis of hospital episode statistics 1991-2002. BMJ. 2004;329:825.

20. Gibbs JL, Monro JL, Cunningham D, Rickards A. Survival after surgery or therapeutic catheterisation for congenital heart disease in children in the United Kingdom: analysis of the central cardiac audit database for 2000-1. BMJ. 2004;328:611.

21. Beral V, Bull D, Doll R, Peto R, Reeves G. Breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83,000 women with breast cancer from 16 countries. Lancet. 2004;363:1007-16.

22. Tate AR, Calderwood L, Dezateux C, Joshi H. Mother's consent to linkage of survey data with her child's birth records in a multi-ethnic national cohort study. Int.J Epidemiol. 2006;35:294-8.

23. Donnan PT, Wei L, Steinke DT, Phillips G, Clarke R, Noone A et al. Presence of bacteriuria caused by trimethoprim resistant bacteria in patients prescribed antibiotics: multilevel model with practice and individual patient data. BMJ. 2004;328:1297.

24. Hypponen E, Davey SG, Shepherd P, Power C. An intergenerational and lifecourse study of health and mortality risk in parents of the 1958 birth cohort: (II) mortality rates and study representativeness. Public Health. 2005;119:608-15.

25. Hardy JR, Leaderer BP, Holford TR, Hall GC, Bracken MB. Safety of medications prescribed before and during early pregnancy in a cohort of 81,975 mothers from the UK General Practice Research Database. Pharmacoepidemiol.Drug Saf. 2006;15:555-64.

26. Hardy JR, Holford TR, Hall GC, Bracken MB. Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database. Pharmacoepidemiol.Drug Saf. 2004;13:749-59.

27. McKeever TM, Lewis SA, Smith C, Collins J, Heatlie H, Frischer M et al. Siblings, multiple births, and the incidence of allergic disease: a birth cohort study using the West Midlands general practice research database. Thorax. 2001;56:758-62.

28. Skytthe A, Kyvik K, Holm NV, Vaupel JW, Christensen K. The Danish Twin Registry: 127 birth cohorts of twins. Twin.Res. 2002;5:352-7.

29. Croft ML, Read AW, de KN, Hansen J, Kurinczuk JJ. Population based ascertainment of twins and their siblings, born in Western Australia 1980 to 1992, through the construction and validation of a maternally linked database of siblings. Twin. Res. 2002;5:317-23.

30. Medlund P, Cederlof R, Floderus-Myrhed B, Friberg L, Sorensen S. A new Swedish twin registry containing environmental and medical base line data from about 14,000 same-sexed pairs born 1926-58. Acta Med.Scand.Suppl. 1976;600:1-111.:1-111.

31. Stazi MA, Cotichini R, Patriarca V, Brescianini S, Fagnani C, D'Ippolito C et al. The Italian Twin Project: from the personal identification number to a national twin registry. Twin.Res. 2002;5:382-6.

32. Smith GC,.White IR. Predicting the risk for sudden infant death syndrome from obstetric characteristics: a retrospective cohort study of 505,011 live births. Pediatrics. 2006;117:60-6.

33. Smith GC, Shah I, White IR, Pell JP, Dobbie R. Previous preeclampsia, preterm delivery, and delivery of a small for gestational age infant and the risk of unexplained stillbirth in the second pregnancy: a retrospective cohort study, Scotland, 1992-2001. Am.J.Epidemiol. 2007;165:194-202.

34. Lawlor DA, Sterne JA, Tynelius P, Davey SG, Rasmussen F. Association of childhood socioeconomic position with cause-specific mortality in a prospective record linkage study of 1,839,384 individuals. Am.J Epidemiol. 2006;164:907-15.

35. Doyle P. The U.K. Human Fertilisation and Embryology Authority. How it has contributed to the evaluation of assisted reproduction

technology. Int.J.Technol.Assess.Health Care. 1999;15:3-10.

36. Chuang CH, Doyle P, Wang JD, Chang PJ, Lai JN, Chen PC. Herbal medicines used during the first trimester and major congenital malformations: an analysis of data from a pregnancy cohort study. Drug Saf. 2006;29:537-48.
37. Cooper WO, Hernandez-Diaz S, Arbogast PG, Dudley JA, Dyer S, Gideon PS et al. Major congenital malformations after first-trimester exposure to ACE inhibitors. N.Engl.J Med. 2006;354:2443-51.
38. Nyboe AA,.Erb K. Register data on Assisted Reproductive Technology (ART) in Europe including a detailed description of ART in Denmark. Int.J.Androl. 2006;29:12-6.
39. Zegers-Hochschild F, Nygren KG, Adamson GD, de MJ, Lancaster P, Mansour R et al. The International Committee Monitoring Assisted Reproductive Technologies (ICMART) glossary on ART terminology. Fertil.Steril. 2006;86:16-9.
40. Nygren KG. The Swedish experience of assisted reproductive technologies surveillance. World Health Organisation, 2006.
41. Koivurova S, Hartikainen AL, Gissler M, Hemminki E, Sovio U, Jarvelin MR. Neonatal outcome and congenital malformations in children born after in-vitro fertilization. Hum. Reprod. 2002;17:1391-8.
42. Fair M,.Cyr M. The Canadian Birth Data Base: a new research tool to study reproductive outcomes. Health Rep. 1993;5:281-90.
43. Wright VC, Chang J, Jeng G, Macaluso M. Assisted reproductive technology surveillance--United States, 2003. MMWR Surveill Summ. 2006;55:1-22.
44. Leunens L, Celestin-Westreich S, Bonduelle M, Liebaers I, Ponjaert-Kristoffersen I. Cognitive and motor development of 8-year-old children born after ICSI compared to spontaneously conceived children. Hum.Reprod. 2006;21:2922-9.
45. Bergh T, Ericson A, Hillensjo T, Nygren KG, Wennerholm UB. Deliveries and children born after in-vitro fertilisation in Sweden 1982-95: a retrospective cohort study. Lancet. 1999;354:1579-85.
46. Gissler M, Louhiala P, Hemminki E. Nordic Medical Birth Registers in epidemiological research. Eur.J.Epidemiol. 1997;13:169-75.
47. Kallen B, Finnstrom O, Nygren KG, Otterblad OP, Wennerholm UB. In vitro fertilisation in Sweden: obstetric characteristics, maternal morbidity and mortality. BJOG. 2005;112:1529-35.
48. Kallen B, Finnstrom O, Nygren KG, Olausson PO. In vitro fertilization (IVF) in Sweden: infant outcome after different IVF fertilization methods. Fertil.Steril. 2005;84:611-7.
49. Kallen B, Finnstrom O, Nygren KG, Olausson PO. In vitro fertilization in Sweden: child morbidity including cancer risk. Fertil.Steril. 2005;84:605-10.
50. Kallen B, Finnstrom O, Nygren KG, Olausson PO. In vitro fertilization (IVF) in Sweden: risk for congenital malformations after different IVF methods. Birth Defects Res.A Clin Mol.Teratol. 2005;73:162-9.
51. Pinborg A, Loft A, Schmidt L, Langhoff-Roos J, Andersen AN. Maternal risks and perinatal outcome in a Danish national cohort of 1005 twin pregnancies: the role of in vitro fertilization. Acta Obstet.Gynecol.Scand. 2004;83:75-84.
52. Pinborg A, Loft A, Nyboe AA. Neonatal outcome in a Danish national cohort of 8602 children born after in vitro fertilization or intracytoplasmic sperm injection: the role of twin pregnancy. Acta Obstet.Gynecol.Scand. 2004;83:1071-8.
53. Pinborg A, Loft A, Rasmussen S, Schmidt L, Langhoff-Roos J, Greisen G et al. Neonatal outcome in a Danish national cohort of 3438 IVF/ICSI and 10,362 non-IVF/ICSI twins born between 1995 and 2000. Hum.Reprod. 2004;19:435-41.
54. Lidegaard O, Pinborg A, Andersen AN. Imprinting disorders after assisted reproductive technologies. Curr.Opin.Obstet.Gynecol. 2006;18:293-6.

55. Lidegaard O, Pinborg A, Andersen AN. Imprinting diseases and IVF: Danish National IVF cohort study. Hum.Reprod. 2005;20:950-4.

56. Hvidtjorn D, Grove J, Schendel DE, Vaeth M, Ernst E, Nielsen LF et al. Cerebral palsy among children born after in vitro fertilization: the role of preterm delivery--a population-based, cohort study. Pediatrics. 2006;118:475-82.

57. Pinborg A, Loft A, Schmidt L, Greisen G, Rasmussen S, Andersen AN. Neurological sequelae in twins born after assisted conception: controlled national cohort study. BMJ. 2004;329:311.

58. Pinborg A, Loft A, Rasmussen S, Nyboe AA. Hospital care utilization of IVF/ICSI twins followed until 2-7 years of age: a controlled Danish national cohort study. Hum.Reprod. 2004;19:2529-36.

59. Mosgaard BJ, Lidegaard O, Kjaer SK, Schou G, Andersen AN. Infertility, fertility drugs, and invasive ovarian cancer: a case-control study. Fertil.Steril. 1997;67:1005-12.

60. Winther JF, Sankila R, Boice JD, Tulinius H, Bautz A, Barlow L et al. Cancer in siblings of children with cancer in the Nordic countries: a population-based cohort study. Lancet. 2001;358:711-7.

61. Gissler M, Jarvelin MR, Hemminki E. Comparison between research data and routinely collected register data for studying childhood health. Eur.J.Epidemiol. 2000;16:59-66.

62. Gissler M, Teperi J, Hemminki E, Merilainen J. Data quality after restructuring a national medical registry. Scand.J.Soc.Med. 1995;23:75-80.

63. Gissler M,.Shelley J. Quality of data on subsequent events in a routine Medical Birth Register. Med.Inform.Internet.Med. 2002;27:33-8.

64. Munk-Jorgensen P, Kastrup M, Mortensen PB. The Danish psychiatric register as a tool in epidemiology. Acta Psychiatr.Scand.Suppl. 1993;370:27-32.:27-32.

65. Libby G, Smith A, McEwan NF, Chien PF, Greene SA, Forsyth JS et al. The Walker Project: a longitudinal study of 48,000 children born 1952-1966 (aged 36-50 years in 2002) and their families. Paediatr.Perinat.Epidemiol. 2004;18:302-12.

66. Hippisley-Cox J, Stables D, Pringle M. QRESEARCH: a new general practice database for research. Inform.Prim.Care. 2004;12:49-50.

67. Jick SS, Kaye JA, Vasilakis-Scaramozza C, Garcia Rodriguez LA, Ruigomez A, Meier CR et al. Validity of the general practice research database. Pharmacotherapy. 2003;23:686-9.

68. Carey IM, Cook DG, De WS, Bremner SA, Richards N, Caine S et al. Implications of the problem orientated medical record (POMR) for research using electronic GP databases: a comparison of the Doctors Independent Network Database (DIN) and the General Practice Research Database (GPRD). BMC.Fam.Pract. 2003;4:14.:14.

69. De WS, Carey IM, Bremner SA, Richards N, Hilton SR, Strachan DP et al. A comparison of the recording of 30 common childhood conditions in the Doctor's Independent Network and General Practice Research Databases. Health Stat.Q. 2004;21-31.

70. Hippisley-Cox J, O'Hanlon S, Coupland C. Association of deprivation, ethnicity, and sex with quality indicators for diabetes: population based survey of 53,000 patients in primary care. BMJ. 2004;329:1267-9.

71. Hippisley-Cox J, Coupland C, Logan R. Risk of adverse gastrointestinal outcomes in patients taking cyclo-oxygenase-2 inhibitors or conventional non-steroidal anti-inflammatory drugs: population based nested case-control analysis. BMJ. 2005;331:1310-6.

72. Hippisley-Cox J,.Coupland C. Risk of myocardial infarction in patients taking cyclo-oxygenase-2 inhibitors or conventional non-steroidal anti-inflammatory drugs: population based nested case-control analysis. BMJ. 2005;330:1366.

73. Hippisley-Cox J,.Coupland C. Effect of statins on the mortality of patients with ischaemic heart disease: population based cohort study

with nested case-control analysis. Heart. 2006;92:752-8.

74. Hippisley-Cox J, Pringle M, Crown N, Coupland C. A case-control study on the effect of hormone replacement therapy on ischaemic heart disease. Br.J.Gen.Pract. 2003;53:191-6.

75. Hippisley-Cox J, Smith S, Smith G, Porter A, Heaps M, Holland R et al. QFLU: new influenza monitoring in UK primary care to support pandemic influenza planning. Euro.Surveill. 2006;11:E060622.

76. Lawrenson R, Wyndaele JJ, Vlachonikolis I, Farmer C, Glickman S. A UK general practice database study of prevalence and mortality of people with neural tube defects. Clin Rehabil. 2000;14:627-30.

77. Births in Great Britain resulting from assisted conception, 1978-87. MRC Working Party on Children Conceived by In Vitro Fertilisation. BMJ. 1990;300:1229-33.

78. Templeton A, Morris JK, Parslow W. Factors that affect outcome of in-vitro fertilisation treatment. Lancet. 1996;348:1402-6.

79. Bradbury BD,.Jick H. In vitro fertilization and childhood retinoblastoma. Br.J.Clin Pharmacol. 2004;58:209-11.

80. Bonduelle M, Wennerholm UB, Loft A, Tarlatzis BC, Peters C, Henriet S et al. A multi-centre cohort study of the physical health of 5-year-old children conceived after intracytoplasmic sperm injection, in vitro fertilization and natural conception. Hum.Reprod. 2005;20:413-9.

81. Ponjaert-Kristoffersen I, Bonduelle M, Barnes J, Nekkebroeck J, Loft A, Wennerholm UB et al. International collaborative study of intracytoplasmic sperm injection-conceived, in vitro fertilization-conceived, and naturally conceived 5-year-old child outcomes: cognitive and motor assessments. Pediatrics. 2005;115: e283-e289.

82. Sutcliffe AG, Barnes J, Wennerholm UB, Loft A, Tarlatzis BC, Ponjaert-Kristoferson I et al. Laterality in five-year-olds conceived by intracytoplasmic sperm injection, standard in vitro fertilisation and natural conception: a European study. BJOG. 2005;112:1397-401.

83. Sutcliffe AG, Bonduelle M, Taylor BW. Major birth defects after assisted reproduction. N.Engl. J.Med. 2002;347:1449-51.

84. Smith GC, Shah I, Pell JP, Crossley JA, Dobbie R. Maternal obesity in early pregnancy and risk of spontaneous and elective preterm deliveries: a retrospective cohort study. Am.J.Public Health. 2007;97:157-62.

85. Smith GC, Shah I, White IR, Pell JP, Crossley JA, Dobbie R. Maternal and biochemical predictors of spontaneous preterm birth among nulliparous women: a systematic analysis in relation to the degree of prematurity. Int.J.Epidemiol. 2006;35:1169-77.

86. Toledano MB, Hansell AL, Jarup L, Quinn M, Jick S, Elliott P. Temporal trends in orchidopexy, Great Britain, 1992-1998. Environ.Health Perspect. 2003;111:129-32

Annex B - Advisory Group Membership

Professor Ian Diamond (Chair)
Chief Executive
Economic and Social Research Council

Dr Richard Barker
Director General
Association of the British Pharmaceutical Industry

Dr Philip J Burstein
Head, Electronic Healthcare Information Initiative
GlaxoSmithKline R&D Ltd

Mr Robin Clark
Director
National Cancer Research Institute - Informatics Initiative

Professor Janet Darbyshire
Director
UK Clinical Research Network/
Medical Research Council – Clinical Trials Unit

Mr Simon Denegri
Chief Executive
Association of Medical Research Charities

Professor Ian Harvey
Professor of Epidemiology
University of East Anglia (FPHM)

Professor Denise Lievesley
Chief Executive
Information Centre for Health & Social Care

Dr John Parkinson
Group Head, General Practice Research Database
Medicines and Healthcare products Regulatory Agency

Mr Nick Partridge
Chair of INVOLVE
Chief Executive, Terrence Higgins Trust

Dr George Sarna
Assistant Director (Strategy)
Medical Research Council

Professor Martin Severs
Associate Dean (Clinical Practice),
University of Portsmouth/
Connecting for Health

Professor Frank Sullivan
Professor of R&D in General Practice and Primary Care
University of Dundee

Mr Jeremy Thorp
Director of Business Requirements
Connecting for Health

Dr Mark Walport
Director
Wellcome Trust

Professor Simon Wessely
Professor of Epidemiology
Institute of Psychiatry, King's College London

Observers

Professor John Newton
Senior Policy Advisor, Department of Health /
Honorary Professor of Public Health & Epidemiology,
University of Manchester

Mr Marc Taylor
Head of Research Policy and Governance,
Department of Health

Professor John G Williams
Director
Welsh Office of Research & Development for Health & Social Care

Dr Louise Wood
Head of Innovation and Industry R&D Relations,
Department of Health

Annex B - continued...

TERMS OF REFERENCE

The group has two purposes against which its value can be judged:

First, to promote collaboration between the UKCRC and CfH in realising the potential benefits for research of data held both at local and national (spine) level of the national care record service infrastructure.

Second, to be responsible for establishing a joint work programme including pilot programmes and reporting its progress.

Sponsorship

The group is co-sponsored by the DH Director of R&D (Professor Sally Davies) and DH Director of IT Service Implementation (Mr Richard Jeavons).

Work Programme Priorities

To provide the research interest input to the IT strategy refresh programme.

To collaborate with the National Programme for IT's Care Record Development Board (CRDB) as a member of its working group on secondary uses.

To test the research capability of the proposed national care record infrastructure using three pilot projects.

Annex C - Simulation Subgroup Membership

Dr Alan Doyle/ Dr Pat Goodwin
Programme Manager
Wellcome Trust

Dr Liam O'Toole
Chief Executive
UK Clinical Research Collaboration

Dr George Sarna
Assistant Director (Strategy)
Medical Research Council

Mr Jeremy Thorp
Director of Business Requirements
Connecting for Health

Mr Paul Willer
Consultant
Association of the British Pharmaceutical Industry

Dr Louise Wood
Head of Innovation and Industry R&D Relations
Department of Health

TERMS OF REFERENCE

The remit of this group was to oversee the commissioning of the four simulations and the ongoing management and steer of this programme of work. This included a series of simulation leads meetings between the simulation leads and members of the Connecting for Health's Secondary Uses Service development team.

Acknowledgements

Simulation Leads and team members

Professor Carol Dezateux
(Observational Epidemiology team)
Director, MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health

Mr Andy Harris
(Prospective Cohort Tracking team)
Systems Architect, UK Biobank

Dr John Parkinson
(Surveillance team)
Medicines and Healthcare Products Regulatory Agency (GPRD)

Mr Rob Thwaites
(Clinical Trials team)
Director
Health Information Factory, GlaxoSmithKline R&D Ltd

Professor N Chaturvedi
(Clinical Trials team)
National Heart & Lung Institute, Imperial College

Professor Rory Collins (Prospective Cohort Tracking team)
Chief Executive Officer, UK Biobank

Dr Lisa Davenport
(Clinical Trials team)
NHS Trust/PCT Liaison, Global Clinical Operations, Johnson & Johnson

Dr John Logie
(Clinical Trials team)
Assistant Director Worldwide Epidemiology, GlaxoSmithKline R&D Ltd

Mr Steve Mott
(Surveillance team)
Chief Executive, DataPharm Communications Ltd

Professor Catherine Peckham
(Observational Epidemiology team)
MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health

Dr Tim Sprosen
(Prospective Cohort Tracking team)
Chief Scientific Officer, UK Biobank

Mr Steve Walker
(Prospective Cohort Tracking team)
Chief Information Officer, UK Biobank

National Programmes for IT / Information Centre for Health and Social Care

Professor Andrew Morris FRSE
Professor of Diabetic Medicine, University of Dundee

Mr Jeremy Thorp
Director of Business Requirements, Connecting for Health

Dr Jon Fistein
Workstream lead for SUS / Access Control, Demographics and Clinical Engagement, Connecting for Health

Ms Lisa Franklin
SUS Programme Director, Connecting for Health

Mr Simon Heathfield
Requirements Analyst, Secondary Uses Service, Connecting for Health

Mr Jamie Kirk
SUS Service Manager, Information Centre for Health and Social Care

UKCRC Secretariat

Dr Liam O'Toole
Chief Executive of the UKCRC and *Acting Chief Executive* of OSCHR

Ms Ngozi Okwudili-Ince
Programme Manager, UKCRC