The Academy *of* Medical Sciences

Data in Safe Havens

March 2014

Workshop report

The Academy of Medical Sciences

The Academy of Medical Sciences is the independent body in the UK representing the diversity of medical science. Our mission is to promote medical science and its translation into benefits for society. The Academy's elected Fellows are the United Kingdom's leading medical scientists from hospitals, academia, industry and the public service. We work with them to promote excellence, influence policy to improve health and wealth, nurture the next generation of medical researchers, link academia, industry and the NHS, seize international opportunities and encourage dialogue about the medical sciences.

All web references were accessed in July 2014.

Contents

Executive Summary

Background

- Greater access to, and linking of, data from research, administration and healthcare provides opportunities to conduct research to advance science, improve patient outcomes, and inform and enhance public health. In order to realise these research benefits it is important that mechanisms are provided to enable such access and linkage whilst upholding the duty of confidentiality and protecting the data subject's right to privacy. Data safe havens are understood to be environments in which these aims can be delivered.
- This workshop, convened by the Academy of Medical Sciences with support from the Medical Research Council and the Wellcome Trust, explored the current landscape to: establish the degree of agreement about the meaning and adequacy of data safe havens; and identify next steps in their development.
- Agreeing on a single definition of 'data safe haven' will be difficult as there is a wide variety of systems in operation. Whilst they are generally understood to hold data securely and provide a safe environment for data analysis, there is no consistency on whether the safe haven: holds identified or de-identified data; provides access to data on site or remotely; processes data and sends them externally; and provides training and support for data users.

Future considerations

- Currently, there is no central registry of safe havens, which makes it difficult to establish any common framework or consistent governance structures. It will be helpful if there is a catalogue of '*where, who, and under what auspices'* of data safe havens.
- Research is generally seen as a bastion of good practice for data stewardship. The *Information Governance Review* (2013), chaired by Dame Fiona Caldicott, notes how researchers have devised robust solutions to enable access to detailed patient information, while ensuring confidentiality is protected.[1] There should be opportunities for researchers to feed into policy developments on the framework and criteria for accrediting data safe havens to reflect existing best practice and ensure their utility for the full range of data use.
- Having a small number of common systems, with sufficient flexibility, is preferable to having numerous safe havens with disparate characteristics. This will be more cost effective and simplify the communication with users and the general public. There will be great benefits and efficiencies if links can be made between data safe havens from administration, healthcare and research. Further work will be required to: clarify and provide consistency across different legislative and policy frameworks; enhance interoperability and linkage; and ensure greater consistency in the governance and operational structures.

[1] Caldicott F (2013). *Information: To share or not to share? The Information Governance Review*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf

- There are many challenges to providing safety of data, but the risks can be managed in a number of ways, including segregation of sensitive data, minimising movement between safe havens, and robust recording and archiving of data usage and access. Agreed criteria for maintaining data safety and clear penalties for negligence or active breach should be established, for instance through HR procedures and contract or data protection legislation. This should go hand in hand with processes and a culture that facilitates appropriate data stewardship. A sector-wide training and accreditation programme that involves academia, industry and the NHS, and which is directed at individuals and institutions could be explored.
- The system should retain sufficient flexibility to ensure the effectiveness of safe havens. It can be difficult to have very specific research goals so the exact data required for analysis is not always clear at the outset. Furthermore, research questions may alter through the course of the investigation. A rigid system that cannot easily accommodate additional data requests will not only slow the research process, it is likely to drive investigators to request more data than perhaps necessary at the start.
- Public engagement is essential to build trust in data use for research. There should be a clear message, tested with the public, which explains: what safe havens are for and what they are not for; what data are and are not included in a safe haven; what data will be used for and what they will not be used for; who will and will not have access to data; and the degree of risks involved. Due to the variety of systems in operation, it will be difficult to agree on a common, single message about what a safe haven is. It will nonetheless be useful to identify common high level characteristics to convey to the public.
- Further work is required to determine metrics to assess the success of safe havens, as well as ways to quantify harm that has occurred from security breaches.
- Developing and running safe havens are expensive. Ongoing funding will be essential to maintain and ensure their continuing functionality. Funding will need to be strategic and co-ordinated across funders and government to be sustainable.

Introduction

A workshop on ‘Data in Safe Havens’, organised by the Academy of Medical Sciences with support from the Medical Research Council and the Wellcome Trust, was held in London on 24 March 2014. The workshop was chaired by Professor Simon Lovestone FMedSci, Professor of Translational Neuroscience at the University of Oxford, and comprised two sections to: (i) explore the range of safe havens currently in existence; and (ii) establish the degree of agreement about the meaning and adequacy of data safe havens and identify next steps in their development.

A number of initiatives aimed at providing greater access to and linking of data from research, administration and healthcare are currently underway. Access to such data sets will prove invaluable in advancing science and improving patient care and services. However, it is important to balance research opportunities with protection of privacy and confidentiality of data subjects. This timely workshop provided the opportunity to discuss the application of data safe havens to address issues surrounding access to personal data under ever increasing public scrutiny.

Summary of presentations

Overview

The morning session started with scene-setting talks by Professor Harry Hemingway, Professor of Epidemiology and Public Health, University College London and Centre Director of the Farr Institute @ London; and Mr Peter Knight, Deputy Director, Research Information and Intelligence, Department of Health, that are summarised in this section.

What is a 'data safe haven'?

Data safe havens aim for wider and deeper linkages of data sets to improve healthcare and health, whilst upholding the duty of confidentiality and protecting the data subject's rights to privacy. The 2008 *Data Sharing Review* recommended the creation of 'safe havens' and, more recently, Dame Fiona Caldicott's *Information: To share or not to share? The Information Governance Review* (Caldicott Review) called for the establishment of 'accredited safe havens' that comply with a set of data stewardship requirements (see Box 1).[2,3] These requirements include compliance with ISO27001 and the Information Governance Toolkit (IGT).[4,5]

Data safe havens can be summarised as an environment where people, working practice, technology, hardware and software are independently accredited, routinely audited, penetration tested, educated and reviewed. They should obey a set of principles that bind users and collaborators to good practice and enforce sanctions for breaches so that the risks of harming research participants are reduced, good research is supported, public trust is upheld and served, and data are processed legally and ethically and treated with respect.

Different types of data safe haven

'Accredited Safe Havens'

The Health and Social Care Information Centre (HSCIC), established under the Health and Social Care Act 2012, is a legally constituted safe haven. Further details about the HSCIC are provided in Annex I but it is approved to transfer data to commissioning organisation Accredited Safe Havens until October 2014.[6] These Accredited Safe Havens are able to process "weakly pseudonymised data" and operate primarily to support care service provision. They are accredited in a three step process of: IGT level 2 compliance; annual

[2] Thomas R & Walport M (2006). *Data Sharing Review*. http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/links/datasharingreview.pdf

[3] Caldicott F (2013). *Information: To share or not to share? The Information Governance Review*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf

[4] ISO27001 is the international best practice standard for an Information Security Management System. See: http://www.27000.org/index.htm

[5] The Information Governance Toolkit is an online system which allows NHS organisations and partners to assess themselves against Department of Health Information Governance policies and standards. See: https://www.igt.hscic.gov.uk/

[6] http://www.hscic.gov.uk/article/3697/Register-of-Stage-One-Accredited-Safe-Havens

audit; and signing of a Data Sharing Contract.[7] There have been concerns, however, that IGT compliance is essentially a self assessment tool with no independent review.

Box 1

The 2008 ***Data Sharing Review*** recommended the development of safe havens as "an environment for population-based research and statistical analysis in which the risk of identifying individuals is minimised".[8] It also called for "a system of approving or accrediting researchers who meet the relevant criteria to work within those safe havens" and that this will require legislation to "ensure that researchers working in 'safe havens' are bound by a strict code, preventing disclosure of any personally identifying information, and providing criminal sanctions in case of breach of confidentiality".

Dame Fiona Caldicott's ***Information Governance Review*** recommended that the "linkage of personal confidential data, which requires a legal basis, or data that has been de-identified, but still carries a high risk that it could be re-identified with reasonable effort, from more than one organisation for any purpose other than direct care should only be done in specialist, well-governed, independently scrutinised and accredited environments called 'accredited safe havens'".[9] The Review report also provides 'Data stewardship requirements for accredited safe havens':

- attributing explicit responsibility for authorising and overseeing the anonymisation process e.g. through a Senior Information Risk Officer;
- appropriate techniques for de-identification of data, the use of 'privacy enhancing technologies' and re-identification risk management;
- the use of 'fair processing notices';
- a published register of data flowing into or out of the safe haven including a register of all data sets held;
- robust governance arrangements that include, but are not limited to, policies on ethics, technical competence, publication, limited disclosure/access, regular review process and a business continuity plan including disaster recovery;
- clear conditions for hosting researchers and other investigators who wish to use the safe haven;
- clear operational control including human resources procedures for information governance, use of role-based access controls, confidentiality clauses in job descriptions, effective education and training and contracts;
- achieving a standard for information security commensurate with ISO27001 and the Information Governance Toolkit;
- clear policies for the proportionate use of data including competency at undertaking privacy impact assessments and risk and benefit analysis;
- standards that are auditable;
- a standard template for data sharing agreements and other contracts that conforms to legal and statutory processes.

[7] http://www.hscic.gov.uk/media/12203/Accredited-Safe-Haven-Accreditation-Process-Stage-1---June-2013/pdf/safe-haven-accred-proc-stage-1.pdf

[8] Thomas R & Walport M (2006). *Data Sharing Review.* http://www.connectingforhealth.nhs.uk/systemsandservices/infogov/links/datasharingreview.pdf

[9] Caldicott F (2013). *Information: To share or not to share? The Information Governance Review.* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf

'Research Safe Havens'

There are a number of established data safe havens that hold anonymous and pseudonymous data sets that support research, such as the Scottish Informatics Programme (SHIP), the Secure Anonymised Information Linkage (SAIL) in Wales and the UK Secure eResearch Platform (UK SeRP). The Government, Medical Research Council, Economic and Social Research Council and National Institute for Health Research have also been investing in a number of data systems for research purposes such as the Clinical Practice Research Datalink (CPRD); the Farr Institute and its component Health Informatics Research Centres; the Administrative Data Research Centres and the Medical Bioinformatics Centres.

Research safe havens generally comply with broad guidelines set out in the *Data Sharing Review* with some demonstrating IGT compliance or holding ISO certification. Research is generally seen as a bastion of good practice for data stewardship and the Caldicott Review notes how researchers have devised robust solutions to enable access to detailed patient information, while ensuring confidentiality is protected. It is worth noting however, that other than the list of Accredited Safe Havens noted above currently, there is no register of data safe havens. This places the onus on collaborators to insist on high standards from each other. Looking to the future, many of the research safe havens may seek Accredited Safe Haven status as well as ISO27001 certification. The Farr Institute, a distributed institute established in 2013 (see Annex II), is now working to facilitate cross-centre collaboration underpinned by consistent governance – with an independent review and audit – across sites.

Establishing data safe havens

The establishment of data safe havens is a lengthy task that may be hindered by the bureaucratic nature of the process, particularly where certifications are sought, and requires oversight by dedicated staff. Technical security implementations, although necessary to protect patient confidentiality, are often not appreciated by users. The main focus, however, should be on people rather than the technical aspects. Outreach, awareness, training and education – particularly in conveying potential identification risks – are absolutely vital to curtail accidental re-identification.

There are a number of people involved in data safe havens, including the senior officer responsible for the overall operation; the officer who bears direct responsibility for the assets (data, hardware, etc); data and service managers that facilitate data access; and users. All these people bear responsibility for risk mitigation but the most risk is likely to exist with the data users, reinforcing the need for safe havens as a secure working environment where accidental disclosure is less likely to occur, deliberate disclosure is harder to achieve, and training and education are provided.

Challenges

There are a number of challenges in the development and operation of safe havens:

- building and maintaining public trust, undermined by recent events such as: the Leveson inquiry into the phone-hacking scandal; uncovering of global surveillance programs by Edward Snowden; and criticisms over the care.data programme (see Annex I) regarding its opaque governance processes and public information campaign;
- complexity of anonymisation processes and their inability to completely prevent the risk of re-identification;[10]
- security breaches – there are widely publicised examples of breaches in the NHS and although we don't tend to hear issues about research, it is not clear if this is due to lack of reporting;
- Data Protection Act 1998 and its interpretation, and the proposed Data Protection Regulation (see Box 2);
- costs associated with the establishment, running and expansion of safe havens – funds are required for technical aspects such as acquisition and maintenance of hardware and software; certification and accreditation; and staffing;
- difficulties in measuring the research impacts and benefits of data safe havens; and
- collaborating at scale and the need to address: consistency across centres; interoperability between heterogeneous data sets; and storage capacity.

Box 2

Since January 2012, the European Union has been developing the **Data Protection Regulation**. The original draft Regulation included a requirement for specific and explicit consent for the use and storage of personal data but provided an exemption for research, subject to certain safeguards. The European Parliament has since adopted amendments that significantly reduce the scope of this research exemption. If they are taken forward, there are potentially significant impacts for health and scientific research across Europe, including the operation of data safe havens. A group of research organisations have been working to ensure that the final Regulation strikes an appropriate balance between the safe and secure use of personal data in research and the rights and interests of individuals.[11]

Models of data safe havens

The second session of the day consisted of a range of presentations on existing data safe havens:

- **Health and Social Care Information Centre** (HSCIC) by Mr Garry Coleman, Head of Data Management Services, Health and Social Care Information Centre (see **Annex I**);
- **Farr Institute** by Professor Ronan Lyons, Professor of Public Health, Swansea University and Centre Director, Farr Institute @ CIPHER (see **Annex II**);

[10] Academy of Medical Sciences (2008). *Personal data for public good: using health information in medical research*. http://www.acmedsci.ac.uk/viewFile/publicationDownloads/Personal.pdf

[11] See for instance the joint briefing, *Protecting health and scientific research in the Data Protection Regulation (2012/0011(COD)): Position of non-commercial research organisations and academics* – April 2014. http://www.acmedsci.ac.uk/viewFile/5363d6423e158.pdf

- **Administrative Data Research Network** (ADRN) by Professor Peter Elias CBE, Warwick Institute for Employment Research, Warwick University (see **Annex III**);
- **UK Data Service Secure Lab** by Ms Melanie Wright, Director of Administrative Data Service and Associate Director of the UK Data Archive (see **Annex IV**);
- **European Medical Information Framework** (EMIF) by Mr Bart Vannieuwenhuyse, European Medical Information Framework Overall Project Coordinator and Senior Director of Health Information Sciences, Janssen R&D (see **Annex V**); and
- **ELIXIR** by Dr Rolf Apweiler, Joint Associate Director and Senior Scientist, EMBL-European Bioinformatics Institute and Advisor for the ELIXIR project (see **Annex VI**).

Summary of discussions

The second part of the meeting consisted of group discussion and feedback. Delegates (see Annex VIII) were asked to consider: the definition and principles of a data safe haven; whether some models of data safe havens are more adequate than others; and key challenges and issues to address for the future development of safe havens and possible solutions. Delegates were asked to consider the full range of data that could be linked and made available, including '-omics', e-health and non-health data sets.

Definition and principles

It was clear from the different models described in the morning session that agreeing on a single definition would be difficult. There was broad agreement that data safe havens hold data securely and provide a safe environment that enables data manipulation. There is no consistency, however, on whether the data held are identified or de-identified or are provided onsite or remotely. There is also no clarity on whether a data safe haven should carry out any of the following activities:

- process and/or 'improve' data on behalf of other entities (which are stored or destroyed after use);
- send (processed) data to the requester; and/or
- provide training and support to the data user.

Delegates agreed that safe havens should enable access to data within a broad framework which reassures users that they are operating within best practice. It was thought that they should share many of the following characteristics:

- timely, reliable and efficient access. It was noted that 'convenience' is one of the most effective ways to prevent misuse;
- a governance system that is appropriate, proportionate, authoritative, trustworthy and ethical;
- a cohort of expertise and provision of training;
- a professional culture with a sense of community among users and a willingness to share;
- assurances and improvements on data quality with appropriate feedback routes;
- clear and precise standards;
- capacity for linkage with external data sets;
- retention of newly created data sets; and/or
- proactive public engagement and increasing public involvement in decision making structures.

There was a suggestion that the term 'safe haven' is not helpful and we should perhaps talk instead of 'models of trusted data usage' for research. A question was also raised about the utility of establishing safe havens under legislation. It was noted that such safe havens may be useful if they were to hold data with legal restrictions for access or potential for harm, although it was acknowledged that the latter will be difficult to define. Variations in legislation across Europe and the differing levels of consent and permissions for data sharing will also add complexity.

Centralised or distributed model

Delegates were broadly in agreement that it will be more helpful to have a small number of common systems – with sufficient flexibility – rather than numerous safe havens with competing characteristics. It was thought that limiting the overall number through a more co-ordinated approach will have several of advantages including:

- focusing resources on developing excellent centres that can be scaled up to European and global level;
- cost containment, particularly as these systems are likely to become increasingly complex and expensive to develop, operate and maintain; and
- simpler messages to convey to users and the general public.

The danger of creating a single 'master safe haven' was, however, highlighted by many of the delegates. It was thought that having a number of safe havens that can undertake a range of data processing activities will be preferable. Greater centralisation was also thought to reduce opportunities for validating data accuracy and quality and has the risk of being perceived as lacking transparency.

Looking at the examples presented in the morning (see Annex I to VI), it was thought that there will be great benefits and efficiencies if links can be made between the HSCIC, Farr Institute and ADRN. The different legislative frameworks for the types of data (e.g. administrative, health) and countries, however, make this difficult. Other areas that need to be addressed include improvements in: interoperability; flow of knowledge; co-ordination of governance structures, data archiving and curation; and funding links.

Mode of data access

Delegates considered the benefits of accessing data held by a safe haven either on site or via remote secure platforms, rather than downloading and removing data for analysis. Not only will this be simpler for governance purposes, it was thought that with an increasing need to incorporate different types of information – such as imaging and '-omics' data – accessing data in this way is likely to become a necessity. The system should not become so rigid, however, as to inhibit any data movement or the ability to combine different data sets.

Having a recognised body that makes decisions on data requests will offer a mechanism to review the quality and appropriateness of analyses conducted with the data. It was also thought to provide a focal point for trust and be beneficial from a public profile perspective. The need for such a body to have a flexible approach, however, was highlighted. It can be difficult to have very specific research goals so investigators are often keen to access a large data set. Furthermore, research questions may alter through the course of the investigation: a rigid review system that cannot easily accommodate additional data requests will not only slow the research process, it is likely to drive investigators to request more data than perhaps necessary at the start.

Risk mitigation

Delegates highlighted that safety is not a binary concept and there are many challenges to providing complete safety of data. Furthermore, there is currently no systematic way to understand the different levels of risks and effectiveness of solutions. Ways to mitigate risks include:

- separation of duties – for instance, de-identification to be only carried out by trusted third parties with robust methods;
- ensuring that log-in details for data access remain tied to an individual. Operating time-bound limits on user access with the possibility of renewal;
- ensuring that each data release is limited to a defined purpose and reviewed by a recognised body. As noted already, however, there needs to be flexibility in such a review process;
- minimising data movement between different safe havens. For international sharing, consider an 'embassy' model of access;
- segregation of data according to sensitivity, as well as clear identification of data that need to be stored in safe havens;
- regular, evidence-based technical 'stress' tests to check for system vulnerability;
- development of detailed mitigation plans against potential threats to safety; and
- comprehensive record of data that have been accessed (by whom) and archived.

Advances in technology have the potential to enhance security, for instance, though development of safer storage locations and improved encryption against re-identification. Some of the examples provided include:

- Data SHIELD (Data Aggregation Through Anonymous Summary-statistics from Harmonized Individual levEL Databases), where distributed computing and parallel analysis are used to enable full pooled meta-analysis of individual level data without accessing individual level data that remain securely stored within each data owner's repository;
- encrypted data sets, which only allow authorised researchers to decode the data ensuring data are protected from unauthorised use; and
- synthetic data sets, which serve as surrogates to the original data thereby ensuring patient confidentiality.

Delegates, however, cautioned against overprotection that may impede linkage between data sets and future data re-use.

The importance of clear penalties against misuse was highlighted, which may also act as a reassurance to the public. It was thought that this may be better at the individual level with institutional penalties being neither practical nor proportionate. Such sanctions may be introduced through HR procedures, contract legislation or data protection legislations. Whilst the fear of being 'struck off' can be a strong individual deterrent to data misuse, delegates stressed the importance of rules and penalties that go hand in hand with processes and a culture that facilitates appropriate data stewardship. As highlighted in the morning presentations, some safe havens are already involved in researcher training and accreditation in the use of data. Delegates thought that there may be merit in a sector-wide training and accreditation programme that involves academia, industry and the NHS. Such sector-wide accreditation may also be directed at the institutional level.

Public engagement

Public trust is crucial for the operation of data safe havens. Delegates noted how long it takes to build public confidence and how easily this can be lost, sometimes through misinformation.

It was thought that a clear message across all safe havens will be helpful, with examples that people can understand. This statement would outline: what safe havens are for as well as what they are *not* for; what data are and are *not* included in a safe haven; what the data will be used for as well as what they will *not* be used for; and who will and will *not* have access to data. Due to the variety of systems in operation, it will be difficult to agree on a common, single message about what a safe haven is. It will nonetheless be useful to identify common high level characteristics to convey to the public; more detailed information can then be tailored to the different safe havens in existence. The statement need not provide in-depth detail about how safe havens are governed but there was a sense amongst delegates that there should be greater transparency about the risks involved including, for instance, through deductive disclosure.

Delegates thought that the messages and terminology used should be tested with the public. It was noted that sometimes it can be difficult to make a clear distinction between using data for research and other purposes (e.g. audit, evaluation, quality assurance, decision making, and marketing), which makes the development of clear messages – as well as different rules of access based on the purpose of data use – problematic. Another issue highlighted was how to assess the success of data safe havens and communicate this to the users and public. What are the most appropriate metrics to be used? Should they include number of outputs, data quality and number of security breaches? In the latter case, is it possible to quantify the actual harm that has occurred from inappropriate disclosure?

There was a suggestion that despite efforts over the years, the debate around benefits and harm of data (re)use is not reaching the public. A more fundamental, open deliberative pubic dialogue involving patient organisations and spanning a number of years may be required.

Sustainability and capacity building

Development of new data safe havens must consider sustainability. Some delegates thought that they should aim for 10-20 year utility with the potential for scaling up, and be able to accommodate the future impact of distributed computing and different types of data.

Ongoing funding will be essential to maintain and ensure continuing functionality of safe havens. It will need to be strategic and co-ordinated across funders and government to be sustainable. Delegates noted that the real costs are associated with human resources and training rather than setting up the infrastructure. In the future, most safe havens may need to operate on a ‘cost-recovery’ or ‘for-profit’ basis with research funders having to consider how to recoup the cost of data usage.

The importance of bioinformatics training, as a specialist course and to improve the generic competencies of researchers, was highlighted. The lack of career structures in research, however, was recognised as a disincentive to attract and retain individuals with data analysis skills.

Concluding remarks from the Chair

The Chair summarised the meeting by reiterating the complexity of this area and importance of treating data with utmost respect. Public engagement is essential for the operation of existing safe havens and development of new systems, although considerable thought is required to ensure its effectiveness. Sustainability will be an ongoing issue: whatever direction we take, it will be expensive and an add-on cost.

In order to facilitate the development of safe havens, the Chair noted that further work will be required on: a common framework (as opposed to legislation, which will take too long); training and accreditation of researchers; proportionate sanctions in cases of breach; acceptable technical solutions to enhance access and security; and how to develop usable and affordable systems.

One immediate action may be to establish a catalogue that lists safe havens and databases that are currently in existence. Such a resource could provide users with a reference and allow assessment of core principles for the development of any common framework and governance structure. It could also act as a central point of information for the public and a basis for public dialogue and engagement activities.

Annex I – Health and Social Care Information Centre (HSCIC), Mr Garry Coleman

The Health and Social Care Information Centre (HSCIC) was established on 01 April 2013. Under the Health and Social Care Act 2012, where directed by the Secretary of State or (for example) NHS England, it has a legal basis to require providers and commissioners of NHS and Social Care to provide data to the HSCIC. Its role includes to:

- set standards that protect patients' confidential information, reduce bureaucracy, improve data quality, and promote system interoperability and standardisation thereby building confidence in the use of information;
- manage essential technology systems and services that support the health and care system; and
- collect, analyse and publish national data and statistical information that helps inform decision making.

care.data

The HSCIC has been commissioned by NHS England to perform the linkage of primary care data and Hospital Episode Statistics (HES, secondary care data) under the care.data programme. The General Practice Extraction Service (GPES) will be the default system by which primary care data from GP practices will be collected and sent to the HSCIC. This has been reviewed and approved by the GPES Independent Advisory Group (IAG) subject to: primary care data being exclusively linked to HES data and no other data sets at this time; these data being collected and used solely for commissioning purposes; and data only being released in anonymised or pseudonymised form. Data extraction was due to start in April 2014 but in light of public concerns, the roll-out was paused to allow time for further consultation.[12]

The data items to be included in the extract have been considered by a clinical informatics expert group with representatives from the British Medical Association and Royal College of General Practitioners. They include: limited patient details (date of birth, postcode, NHS number and gender), events (vaccinations, diagnoses, biological values and all NHS prescriptions), and referrals (date and reason). The primary care data set will only be collected for the last rolling four months from the date of extract.[13]

Safe havens: personal perspective

Full transparency of process, purpose, sharing and use *in conjunction with* proactive engagement with the public is essential. Having appropriate standards and audits are also important. These should enable data re-use, although this should not be seen as a right:

[12] Subsequent to the meeting, the care.data programme has confirmed it will take a phased approach to implementation. It will work with between 100 and 500 GP practices, so-called 'pathfinders', to test, evaluate and refine all aspects of the data collection process ahead of national roll-out. A key part of the pathfinders will be an evaluation against agreed objectives and success criteria, resulting in a decision as to whether to proceed to the point of extracting the data. The pathfinder practices are currently being selected and it is expected that the decision on whether the pathfinder sites are ready to move to the extraction stage will take place in autumn 2014.

[13] This proposal will be discussed further with GPs and Clinical Commissioning Groups as part of the pathfinder stage to determine if data should be collected for an earlier period of time.

access should only be granted in the correct circumstances, set out in published policies/rules, and users should be mindful and respectful of the data they wish to use.

Annex II – Farr Institute, Professor Ronan Lyons

The Farr Institute of Health Informatics Research was established in 2013 as a distributed research institute comprising four nodes spread across the UK (led from University College London, the University of Manchester, Swansea University, and the University of Dundee) with funding from a 10-funder consortium, including the Medical Research Council. Its aims are to:

- create a physical and electronic infrastructure to support and accelerate the Centres' collaborative work;
- support partnership by providing a physical structure to co-locate NHS organisations, industry, and other UK academic centres;
- facilitate collaboration, sharing of data sets, and adoption of common standards; and
- develop new opportunities for future linkage and analysis of data at scale.

To ensure patient confidentiality, there is an increasing trend to separate out: identity matching and first stage pseudonymisation; data repositories and hosted analysis platforms. Identity matching and first stage pseudonymisation should be undertaken by a separate organisation to that holding de-identified data or, if this is not possible, by a managerially distinct part of the same organisation; to this end, a national identity indexing service with the sole task of performing identity matching would be of value in the UK. Data repositories ideally should only hold pseudonymised data and not take analysis of raw data sets. The hosted analysis platforms should run only approved projects carried out by approved researchers who use curtailed data sets containing the minimum number of variables required for analysis. Researcher output should be scrutinised for disclosure risk before release.

The Farr itself works with a number of safe havens including:

- safe havens operated by, or in partnership with, NHS organisations such as Secure Anonymised Information Linkage (SAIL) or the Electronic Data Research and Innovation Service (eDRIS);
- safe havens being developed by, or accredited by, the Health and Social Care Information Centre;
- UK Secure eResearch Platform (UK SeRP), an MRC funded development which allows remote access to linked de-indentified data from administrative data, trials and cohorts; and
- other platforms currently in development.

Public engagement is necessary to instil trust in research using patient data, although there are challenges in explaining complex concepts in a concise, easy-to-understand fashion. A variety of approaches are being trialled in the UK and initiatives, such as the Consumer Panel for Data Linkage at Swansea University, are aimed at increasing public involvement in decision making structures. Other Farr Institute initiatives include Public Engagement and Innovative Governance work streams. The latter seeks to promote best practice in governance for research safe havens bearing in mind the trade-off between harms, benefits and utility of different systems for privacy protecting data linkage.

Annex III – Administrative Data Research Network (ADRN), Professor Peter Elias CBE

The Administrative Data Research Network (ADRN) was established following a review by the Administrative Data Taskforce, which commissioned expert group reports into models of data access and linkage, public engagement, and legal and ethical issues.[14] The taskforce recommended that:

- an Administrative Data Research Centre (ADRC) should be established in each of the four countries in the UK. These would be safe settings where data linkage would be undertaken and made available for analysis;
- legislation should be enacted to facilitate research access to administrative data and to allow data linkage between departments to take place more efficiently;
- a single UK-wide researcher accreditation process, built on best national and international practice, should be established;
- a strategy for engaging with the public should be instituted, encompassing procedures for raising public awareness about the need for such research and involving them in decision-making regarding the administrative data to be accessed and linked; and
- sufficient funds should be put in place to support improved research access to and linkage between administrative data to support the various activities associated with the previous recommendations.

Following a successful bid for funding in 2012, the ADRN has been setting up the partnership arrangements with the relevant statistical authorities, commissioning an ADRC in each country of the UK and an overarching coordinating Administrative Data Service (ADS), and establishing a governing body for the ADRN (the ADRN Board), to guide the strategic direction of the Network and report annually to the UK Statistical Authority. The ADRN will report directly to Parliament, rather than Government departments that supply the data.

The Economic and Social Research Council defines a safe haven as a database that does not hold any personal identifiers (trusted third parties are responsible for holding such data sets and for data linkage), which can only be accessed by accredited researchers for use on site, by remote submission or remote terminals.

Remaining tasks for the ADRN include: public engagement and communication (getting the right messages to opinion formers), establishing links to the Farr Institute (ADRCs are often in close proximity to the four nodes of the Farr), providing input to help shape the legal environment, engaging with private sector interests, and promoting the ADRN as an international resource (although this will require caution since it will have to adhere to international legal frameworks).

[14] Administrative Data Taskforce (2012). *The UK Administrative Research Network: Improving Access for Research and Policy.* http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf

Annex IV – UK Data Service Secure Lab, Ms Melanie Wright

Safe havens: personal perspective

A particular variety of data safe haven is a research data centre, which provides a virtual environment for accessing and analysing data held on a secure central server using an encrypted connection, disconnected from a user's local PC, local network and internet. The principle is that access is distributed, but data stay in one place. This place should provide a 'home from home' environment with familiar software tools and collaborative working areas to incentivise use. A data safe haven should also have a role in project and research approval, training of users, formulation of user agreements, standards-based statistical disclosure control, an effective penalties policy, and researcher management. Five criteria should be adopted to ensure data security:

1. safe data: anonymisation/de-identification to ensure data safety;
2. safe project: valid statistical purpose to ensure project safety;
3. safe place: technical controls around data to provide a safe setting;
4. safe output: disclosure control of results to ensure output safety; and
5. safe people: trusted, trained researchers to ensure people safety.

The highest security risks lie with the researchers, one of the only elements that cannot be 'controlled' by the safe haven and who, by human error or with intent, may breach data protection policies. Certain safeguards can be put in place to counter these risks including licensing, training, enforcing breaches penalties, advocating community self-policing, and evidence-based risk management. Making the data safe haven as convenient and user-friendly as possible is also one of the most effective positive security measures.

UK Data Service Secure Lab

The Secure Lab was launched in March 2011 and contains 45 data sets which were previously unavailable or only available from an onsite facility (no remote access). Almost all of these data sets contain 'sensitive' variables such as date of birth, postcode/grid-references, detailed financial data of business organisations, and health indicators such as alcohol consumption, sexual orientation, and linked school report data.

Applications for data use are vetted by data owners/data access committees and the safe haven is audited annually to ensure that robust access procedures, tested secure storage methods, procedures for processing, handling data and for managing statistical output requests, and methods of dealing with security events are in place. Internally, an Information Security Management Group is responsible for ensuring archive complies with ISO, Government security, and audit requirements. Staff is also very much engaged and training/continuous improvement is provided.

The experience of the Secure Lab so far has been robust with only a small number of easily managed minor incidents. The need for greater transparency and understanding of data owner needs from the offset is imperative. Data safe havens should be a collaborative effort between service, data owners and researchers for most effective results – cultures of suspicion do not breed best behaviour, whereas ownership and community spirit do.

Although challenges remain, such as resources and efficient working practices, it has been a success story so far: there have been no data breaches, high impact research has been undertaken, and research and service communities have been fostered.

Annex V – European Medical Information Framework (EMIF), Mr Bart Vannieuwenhuyse

European Medical Information Framework (EMIF) is part of the Innovative Medicines Initiative (IMI), Europe's largest public-private partnership aimed at accelerating the development of medicines. The pharmaceutical industry, represented by the European Federation of Pharmaceutical Industries and Associations (EFPIA), and the European Union, represented by the European Commission, have jointly pledged €2 billion in funds towards IMI projects.

More specifically, the EMIF project is working towards creating a framework that enables efficient access and re-use of existing human health data across Europe, in view of determining biomarkers of Alzheimer's disease and metabolic complications in obesity. It engages 56 partners across 14 European countries and aims to be the trusted European hub for health care data intelligence. It seeks to capture so-called 'real world data' from clinical practice and patients following product launch in order to inform drug discovery and development. Many types of data are available through the consortium, including data from primary and secondary care, administrative data, and data from registries, cohorts and biobanks. Combined, the framework will hold data of more than 52 million subjects from seven EU countries.

Variations in EU and national policies, as well as differing levels of consent and permissions for data sharing, present major challenges for setting up the platform and require careful administration to ensure privacy regulations are respected. To facilitate this, EMIF is adopting a federated solution in which several self-sustained and functional databases across Europe, each responsible for adhering to privacy legislation and data sharing permissions, can be queried through a single portal.

EMIF members will be able to access a repository of aggregated results for commonly needed data items. For more specialised queries, a transient data pool held by a fully audited trusted third party can be generated. This enables data providers to release only the minimal amount of data needed for a study for a limited period of time. These data releases are enabled by data sharing agreements and are subject to ethical approval and compliance with disclosure and usage policies.

Annex VI – ELIXIR, Dr Rolf Apweiler

ELIXIR is tasked with building a sustainable European infrastructure for biological information to support life science research and its translation into benefits for society, bioindustries, medicine and the environment. ELIXIR is looking to provide access to many types of data, including proteomics, genomics, transcriptomics, and chemistry amongst others, whilst ensuring that they are properly integrated, optimised and adhere to privacy regulations. ELIXIR is also working to develop common standards (format, ontologies and guidelines) and to provide tools (access, search and analysis), compute provision (storage, network and computing), and much-needed training for data sharing and data management across Europe.

ELIXIR has a distributed 'hub-and-node' infrastructure model, with a single hub located at the EMBL-EBI premises in the UK and an increasing number of nodes in research centres across Europe. The hub has responsibility for the overall management and coordination of ELIXIR, as well as delivering services. The nodes research and develop bioinformatic services and training activities, which they deliver through their own 'brands', and are responsible for collaborations with academic and industry partners, who provide and access data.

ELIXIR is currently undertaking several pilot projects aimed at addressing key challenges in biomedical research including: virtual workspaces with the capacity to download large reference data sets; efficient transfer of large data sets between institutes; securing access to and ensuring future capacity for genomic-phenomic data sets; and interoperability and integration of protein and genetic databases for biomedical research.

Future challenges for life science data services lie in the scale and funding of infrastructure to support increasing numbers of data users, interoperability between heterogeneous data sets including clinical and translational data, storage capacity, and virtual research environments whilst ensuring privacy and ethical concerns are respected at all times.

Annex VII – Programme

24 March 2014

Wellcome Trust, Gibbs Building, 215 Euston Road, London, NW1 2BE

09:00 – 09:30	***Registration***
Session 1: Welcome and background	
09:30 – 09:45	Introduction and objectives *Chair, Professor Simon Lovestone FMedSci*
09:45 – 10:00	Wider UK context *Professor Harry Hemingway*
10:00 – 10:15	Update from the Department of Health *Mr Peter Knight*
Session 2: Models of safe haven	
10:15 – 10:30	Health and Social Care Information Centre *Mr Garry Coleman*
10:30 – 10:45	Farr Institute *Professor Ronan Lyons*
10:45 – 11:00	Administrative Data Research Network *Professor Peter Elias CBE*
11:00 – 11:20	Q&A session
11:20 – 11:35	***Tea/coffee***
11:35 – 11:50	UK Data Service Secure Lab *Ms Melanie Wright*
11:50 – 12:05	European Medical Information Framework *Mr Bart Vannieuwenhuyse*
12:05 – 12:20	ELIXIR *Dr Rolf Apweiler*
12:20 – 12:45	Q&A and discussion
12:45 – 13:45	***Lunch***
Session 3: Group discussion	
13:45 – 15:15	Separate into groups to address: • Are some models of data safe havens more adequate than others? • What are the key questions to be addressed? What solutions are required?
15:15 – 15:30	***Tea/coffee***
15:30 – 16:15	Regroup for feedback: • Share points raised: include principles, challenges and implementable ideas for the way forward
16:15 – 16:30	Summing up and next steps *Chair, Professor Simon Lovestone FMedSci*
16:30	***Close***

Annex VIII – Delegate list

Dr Rolf Apweiler, Joint Associate Director and Senior Scientist, EMBL-European Bioinformatics Institute and Advisor for the ELIXIR project

Professor Dame Valerie Beral DBE FRS FMedSci, Director, Cancer Epidemiology Unit, University of Oxford

Professor Martin Bobrow CBE FRS FMedSci, Professor Emeritus, University of Cambridge

Mr Garry Coleman, Head of Data Management Services, Health and Social Care Information Centre

Ms Vanessa Cuthill, ESRC Team Head/Strategic Lead for Administrative Data Research Network, Economic and Social Research Council

Professor Carol Dezateux CBE FMedSci, Professor of Paediatric Epidemiology, University College London

Ms Natasha Dunkley, Confidentiality Advice Manager, Health Research Authority

Professor Peter Elias CBE, Professor at the Warwick Institute for Employment Research, Warwick University

Dr Catherine Elliott, Head of Clinical Research Support and Ethics, Medical Research Council

Mr David Evans, Senior Policy Officer, Strategic Liaison, Information Commissioner's Office

Professor David Ford, Professor of Health Informatics, Swansea University

Dr Robert Frost, Policy Director, Medical Policy, GlaxoSmithKline

Dr Amadou Gaye, Postdoc – DataSHIELD development, University of Bristol

Professor Ruth Gilbert, Deputy Director, Administrative Data Research Centre – England, and Professor of Clinical Epidemiology, University College London

Dr Jane Green, Clinical Epidemiologist, Cancer Epidemiology Unit, University of Oxford

Professor Harry Hemingway, Professor of Epidemiology and Public Health, University College London and Centre Director of the Farr Institute @ London

Dr Kerina Jones, Associate Professor of Health Informatics, College of Medicine, Swansea University

Mr Peter Knight, Deputy Director, Research Information and Intelligence, Department of Health

Ms Katherine Littler, Senior Policy Adviser, The Wellcome Trust

Professor Simon Lovestone FMedSci, Professor of Translational Neuroscience, University of Oxford

Professor Ronan Lyons, Professor of Public Health, Swansea University and Centre Director, Farr Institute @ CIPHER

Baroness Onora O'Neill CH CBE HonFRS FBA FMedSci, Professor Emeritus, University of Cambridge

Dr Dermot O'Reilly, Clinical Senior Lecturer, Queen's University Belfast

Dr John Parkinson, Director, Clinical Practice Research Datalink

Dr Stephen Pavis, Head of Programmes, Scottish Informatics Programme

Dr Nicola Perrin, Head of Policy, The Wellcome Trust

Dr Mark Pitman, Research Programme Manager, Medical Research Council

Dr Mary Rauchenberger, Head of Data Management Systems, MRC Clinical Trials Unit

Dr Bina Rawal, Medical, Innovation & Research Director, ABPI

Professor Martin Richards, Emeritus Professor of Family Research, University of Cambridge
Mr Bart Vannieuwenhuyse, European Medical Information Framework Overall Project Coordinator and Senior Director of Health Information Sciences, Janssen R&D
Mr Phil Walker, Head of Information Governance, Department of Health
Dr Hazel Wardrop, Data Linkage Research Officer, Centre for Longitudinal Studies, Institute of Education
Ms Melanie Wright, Director of Administrative Data Service and Associate Director of the UK Data Archive

<u>Secretariat</u>
Mr Ben Bleasdale, Policy Intern, Academy of Medical Sciences
Dr Claire Cope, Policy Officer, Academy of Medical Sciences
Dr Naho Yamazaki, Head of Policy, Academy of Medical Sciences

Annex IX – Steering group

Professor Martin Bobrow CBE FRS FMedSci, Professor Emeritus, University of Cambridge
Professor Paul Burton, Professor of Infrastructural Epidemiology, University of Bristol
Professor Carol Dezateux CBE FMedSci, Professor of Paediatric Epidemiology, University College London
Mr Peter Knight, Deputy Director, Research Information and Intelligence, Department of Health
Professor Bartha Knoppers, Director of the Centre of Genomics and Policy, McGill University, Canada
Professor Simon Lovestone FMedSci, Professor of Translational Neuroscience, University of Oxford
Baroness Onora O'Neill CH CBE HonFRS FBA FMedSci, Professor Emeritus, University of Cambridge

The Academy *of* Medical Sciences

Academy of Medical Sciences
41 Portland Place
London, W1B 1QH
+44(0)20 3176 2150

info@acmedsci.ac.uk
www.acmedsci.ac.uk

Registered Charity No. 1070618
Registered Company No. 3520281