phgfoundation
making science work for health

Linking and sharing routine health data for research in England

Authors

Dr Charlotte Warren-Gash

Acknowledgements

The author would like to thank the data users who kindly agreed to be interviewed for this report as well as Alison Hall, Dr Mark Kroese and Dr Sobia Raza of the PHG Foundation for helpful comments and discussions.

NB: URLs in this report were correct as at 1 May 2017

This report can be downloaded at:
www.phgfoundation.org

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK

Tel: +44 (0) 1223 761 900

Correspondence to:
charlotte.warren-gash1@lshtm.ac.uk
mark.kroese@phgfoundation.org

How to reference this report:
Linking and sharing routine health data for research in England
PHG Foundation (2017)
ISBN 978-1-907198-27-4

Contents

Executive summary

Healthcare is generating increasingly large amounts of routine data, for example through electronic health records (EHRs) and public health surveillance systems, which provide important opportunities for health research.

Research using routine data is, however, subject to various challenges, which limit its potential to benefit society.

This report sets out the barriers and facilitators to linking and sharing de-identified routine health data between organisations for research in England using the example of EHRs and infection surveillance data.

Objectives of this report

- To conduct a series of semi-structured interviews to understand researchers' experiences of accessing and working with electronic health data on infections in England
- To compare this with models of good practice elsewhere in the UK and internationally
- To make recommendations informed by interviews and the wider literature to improve the experiences of data users and ultimately to facilitate more effective population health research to be conducted using these data sources

Policy recommendations

- Establish systems and incentives to encourage secure data linkage and sharing for research in England
- Increase capacity for data linkage and sharing by public organisations
- Streamline procedures to enable appropriate and efficient access to routine health data for research
- Improve transparency and communication around routine health data access and use between data provider organisations and researchers
- Provide better support for researchers working with routine health data

See page 18 for the detailed policy recommendations.

Health data

Healthcare is generating increasingly large amounts of routine data, which provide important opportunities for health research. Routine data include data from clinical encounters in primary and secondary care, disease registries, vital registrations, public health surveillance and administrative claims databases, among other sources[1]. These datasets offer opportunities to conduct powerful efficient research using novel analytic approaches to inform clinical and public health services[2]. Research using routine data is, however, subject to various challenges – technical, governance, political and financial – that limit the realisation of potential health system benefits[3]. As routine health data are not primarily collected for research, linkage between several different sources may be needed to improve the quality and completeness of a dataset for answering particular health questions. Data linkage and sharing between data providers and end users is often subject to additional barriers, which further limit the utility of these rich datasets for research.

Electronic health records

Electronic health records (EHRs) are digital records of a patient's health and care. They contain information on factors such as demographics, lifestyle, clinical diagnoses, engagement with health services, prescriptions, investigations and medical procedures. EHRs are held at GP surgeries, hospitals, out-patient clinics and other community care settings and vary in format, content, software system, access permissions and use[4]. In England, a range of data providers curate pseudonymised EHR datasets for research including NHS Digital (secondary care data), the Office for National Statistics (vital registration data) and multiple primary care data providers such as the Clinical Practice Research Database (CPRD), The Health Improvement Network (THIN), IMS Health and QResearch.

Infection surveillance data

Infection surveillance in England is coordinated by Public Health England (PHE), which contributes to UK-wide, European and global health protection networks. Data on infections comes from multiple systems including laboratories, clinical reports of statutory notifiable diseases and anonymous syndromic surveillance e.g. of gastrointestinal or influenza-like illnesses in primary care, emergency departments or calls to NHS 111. These datasets are vital to monitor and control infectious disease outbreaks and evaluate health interventions such as vaccination programmes[5]. Data flows for health protection are complex and differences in data collection systems, sharing frameworks and governance structures limit the ability to harmonise infection surveillance data across the devolved nations. Infection data can also be linked to other datasets for public health surveillance or research.

National policy

The importance of optimising EHRs for clinical care and research is recognised in a series of policy documents and initiatives. These include a commitment in the NHS Five Year Forward view to a 'paperless' NHS by 2020[6], with plans outlined further in the National Information Governance Board document *Personalised health and care 2020: Using data and technology to transform outcomes for patients and citizens: a framework for action*[7].

Similarly, the key role of infection surveillance data is described in the Department of Health strategy *Public Health surveillance: Towards a Public Health Surveillance strategy for England*[8]. In a stakeholder consultation to inform this strategy, data linkage and sharing was the most commonly identified priority by stakeholders to strengthen the public health surveillance function of PHE. PHE's vision is set out in the report, *From evidence into action: opportunities to protect and improve the nation's health*[9], which, together with its Strategic Plan for 2016-2020[10], outline the need to improve the capture and use of national infection surveillance data. *The Review on Antimicrobial Resistance*[11], commissioned by the UK Prime Minister in 2014, also highlights the potential for newer big data approaches to improve global surveillance of drug-resistant infections and inform public health responses.

UK research funders also increasingly recognise the potential of interdisciplinary health informatics research using large routinely collected datasets to improve health and care.

National initiatives to promote research using routine health data

General

The Farr Institute is a UK-wide research collaboration comprising 21 academic and health institutions, funded by a consortium led by the Medical Research Council. As well as delivering high quality translational research, the institute aims to develop new infrastructure, technologies and standards for health informatics research, create partnerships and develop researchers' skills for working with complex datasets. The forthcoming establishment of a new multifunder UK Institute for Health and Biomedical Informatics Research will increase strategic support for research in this area.

Specific to infection

The 13 Health Protection Research Units (HPRUs), funded by the National Institute for Health Research, are partnerships between universities and Public Health England designed to undertake leading health protection research. The majority of HPRUs focus on infection research and have galvanised plans for wider linkage between infection surveillance datasets and other data including EHRs.

Interviews

Semi-structured interviews about research using linked EHRs and infection surveillance data were carried out between January and March 2017 with a purposive sample of data users. These ten researchers were from different organisations and ranged in seniority from post-doctoral level to professors. All had experience of using linked EHRs including infection data for research. The interview schedule and more details of methods are given in appendix 1.

Types of research carried out using linked EHRs and infection data

Data users reported carrying out a wide range of observational epidemiological research studies using EHRs linked to or including infection data. From England, they had used primary care data e.g. from the Clinical Practice Research Datalink (CPRD), secondary care data e.g. Hospital Episode Statistics (HES), birth and death registration records from the Office for National Statistics (ONS), pre-entry migrant screening data and disease-specific observational cohorts based on routine clinical data. In addition, they had analysed national infection surveillance datasets on acute respiratory viruses such as influenza and respiratory syncytial virus as well as TB, HIV, bloodstream and healthcare associated infections. Participants also reported conducting infection-related research using similar Scottish linked data sources, insurance claims data from the United States and combined EHR and claims data from a range of European databases.

Participants had used these data sources to:

- Describe infectious burden over time and between countries, regions or healthcare settings
- Answer aetiological questions about risk factors for infections or infectious complications
- Evaluate screening programmes and interventions
- Inform health service planning
- Conduct methodological research

Advantages of using routinely collected EHR data

Data users reported the following benefits to using routine health data for research:

- Saving time and money because there is no expensive and logistically challenging data collection
- Enhancing power to answer some research questions. Linkage between datasets may give a larger sample size than using a single data source, which is especially important for rare conditions and for subtypes of diseases such as dementia or coronary heart disease
- Giving a greater phenotypic depth i.e. more detailed accurate information on the disease under study using information that is obtained from multiple linked datasets

- The ability to recreate the longitudinal pathway of a patient from multiple healthcare datasets, and give repeated measurement of risk factors
- The fact that EHRs represent whole populations rather than just those who have agreed to participate in research. This gives the ability to create historical cohorts of under-represented populations such as migrants, homeless people and intravenous drug users, in whom infection is often an important outcome. It reduces loss to follow up in these groups and allows research into factors affecting disengagement from care

Specific advantages of linked EHRs and infection surveillance data

The advantages of linking EHRs and infection surveillance data for research reported by data users were:

- The enhanced quality, detail, completeness and accuracy of EHR data on infections, compared to infection diagnoses in unlinked EHRs, which are often non-specific and based on clinical rather than microbiological diagnoses
- The fact that some infections cannot be identified at all without linked infection surveillance data e.g. some bloodstream infections, molecular diagnoses of TB or laboratory-confirmed vaccine preventable diseases
- The enhanced range of policy-relevant research questions that can be addressed. While infection surveillance data alone can be used to describe the incidence and prevalence of infections, using infection surveillance data linked to EHRs allows other questions to be investigated such as:
  - Effects of infections on complications or comorbidities
  - Other outcomes of infections
  - Healthcare usage associated with infections
  - Effectiveness of standard therapies such as steroid and antibiotic therapy in conditions such as chronic obstructive pulmonary disease
  - Evaluation of major policy initiatives such as care bundles for sepsis

  Data users noted that addressing these questions will inform clinical treatments, service provision and targeting of prevention initiatives such as programmes to reduce inequalities in vaccine uptake. Optimising use of linked infection surveillance data linked to EHRs might also obviate the need for some of PHE's enhanced public health surveillance practices
- Finally the opportunity to generate hypotheses in a more agnostic way and use a range of sophisticated analytical approaches e.g. machine learning in large linked datasets was noted

Obtaining access to linked EHR-infection surveillance data

Data users reported a range of experiences of obtaining access to infection data. Ease of data access varied depending on whether data are pre-linked; geography of data provider; and researcher status as a university researcher versus an employee of the data provider.

- In general, pre-linked data e.g. from primary care linked to hospital records and ONS mortality data were more straightforward to obtain than data using new linkages, although timescales could still be slow, taking years rather than months
- Obtaining new linked data from Scotland was described as smoother than in England, due to there being a small infection surveillance team who are keen to collaborate with external researchers
- Researchers working within data providers such as PHE either as employees or on secondment from a university with an honorary contract had more success accessing linked EHR-infection surveillance data than researchers based within universities

For new EHR-infection data linkages, the following issues were raised:

- It is time-consuming i.e. it can take several years. This has knock on effects, with dissemination of important research findings being delayed due to funding running out, researchers moving institutions and late publication of results. It can negatively affect PhD student progress and progression for early career researchers
- Although appropriate information governance and data security is essential, the many, often sequential processes needed to access linked data can hinder research. These include obtaining grant funding, data sharing agreements, honorary contracts, confidentiality advisory group approval, ethical approval and other approvals. The more data sources are included, the more complex gaining access becomes because each data source has different access methods (which may be opaque) and requires different permissions. This results in wasted researcher time and disproportionate effort
- It can be frustrating and challenging. In some areas, there was a perceived reluctance of data providers to allow data access to external researchers or other research organisations, perhaps due to issues of competition, control and desire for reciprocity. This meant that some university researchers with grant funding to set up a new data linkage were unable to arrange meetings with data providers. For others, despite meetings and verbal agreements, no data were provided. This led to several projects being abandoned or carried out using data from different countries
- Lack of data linkage capacity can mean that researchers need to go to the data provider to perform the data linkage themselves
- When data access is obtained, this is usually limited to a single purpose. It was noted that accessing linked data for a broader purpose would be more efficient and hypothesis-agnostic (though there are regulatory limits to the breadth of consent that can be given under the Data Protection Act and the forthcoming General Data Protection Regulation[12])

New initiatives were described through the HPRUs such as primary care data being linked to electronic child health record data in some areas. Although primarily intended to enhance clinical support for children, this will also potentially bring new opportunities for research in the future.

Barriers to linked EHR research

Data users described several types of barriers to research using either linked EHRs in general or linked EHRs and infection surveillance data specifically. These were:

Technical: data content and quality

- Linked datasets are complex. Substantial time-consuming work is required to clean, transform and manipulate data. Researchers must be able to handle discordance between datasets. There is a need for good technical, epidemiological and topic expertise to deal with the methodological and interpretational barriers
- Lack of transparency about data linkage processes is a barrier to accurate interpretation of results. Data users also noted that there was a lack of systematic sharing of codes and algorithms for data linkage and cleaning. A lack of re-use of datasets leads to repeated creation of similar linkages, which involves additional data cleaning time and use of resources
- Technical barriers intrinsic to some datasets include a lack of common identifiers such as NHS number to use for data linkage. This was noted for particular populations e.g. migrants and homeless people and in particular settings such as genito-urinary or HIV medicine where, to enhance patient confidentiality, a clinic's records may not link to the main hospital record. It leads to a need for probabilistic linkage, which may require expertise in SQL (Structured Query Language) coding and being able to manipulate largescale datasets within an SQL environment
- For infection surveillance data, changes in the data systems used over time may make it difficult to interpret temporal trends. In some geographic locations, even when linked EHRs and infection surveillance data are available, small numbers can limit the power and generalisability of findings
- Certain research questions cannot be answered using routine data, so to some extent, the research agenda is dictated by the scope of data available

Governance: consent and confidentiality

- Information governance for researchers can be overly complicated and disproportionate to the risks involved in protecting research subjects' data. Understanding and negotiating the legal, ethical and governance frameworks and requirements may be a barrier to data access for researchers unfamiliar with using linked EHRs

- Data users felt that unfounded concerns held by data providers about the legitimacy of data sharing, especially in situations where there was a lack of individual patient consent, were a barrier to research. This was exacerbated by recent problems with the care.data programme undermining public trust. A specific issue with risk averse practices was raised by several data users, who perceived that some data providers lacked the expertise to apply Data Protection Act requirements in a proportionate and workable way
- The importance of dealing effectively with risks of re-identification of individuals when linking several data sources and small numbers in some cells was raised; this has been the subject of litigation in Scotland[13]. Nevertheless it was felt that current ethical and technical requirements for data linkage generally deal effectively with this issue

Cultural: willingness and capacity of data providers to support linked EHR research

- For data providers who hold data for a reason other than research e.g. public health surveillance or direct patient care, facilitating research using linked EHRs is not a core function or priority
- NHS staff using hospital systems may lack the ability and capacity to collect, extract and manipulate data for research
- Other data providers may not be willing to help university researchers use linked EHR datasets for research. Data users believed that some data providers may not see the benefits of linking and sharing data; for others it might be about control and ownership of the data; in other cases long delays in obtaining linked datasets were blamed on a lack of database administrators and data managers with expertise in probabilistic linkage
- Lack of engagement with scientists by data providers is a barrier, e.g. leading to a lack of feedback loops whereby it is impossible for researchers to get data extraction errors corrected
- In some research areas, patient groups may be very involved with research but have negative attitudes towards data sharing. This might be because of worries about confidentiality and ethics as well as a history of stigma, and these attitudes hinder data sharing efforts

The following issues were noted by data users to apply specifically to infection surveillance data:

- There is a lack of transparency about the datasets and access procedures. PHE are the main provider of infection surveillance data in England yet these data are not mentioned in the PHE official data release, so are not available to *bona fide* external researchers. It is unclear who to communicate with and what processes need to be undertaken to access the different datasets
- There is a need for research collaborations with data providers. Access to some datasets is currently based on personal relationships with data providers. Ownership issues may lead to an unwillingness to share data with external researchers, even when the legal basis for data sharing is clear. Data users feel that government organisations may select collaborators whose research is in line with their policy objectives but not in competition with them
- Even if researchers have funding for data linkage work, there are no established models for reimbursing data provider time

Logistics: time and cost

- The long timescales to obtain new linked EHR datasets for research are perceived to lead to a lack of feasibility and difficulty getting grant funding for this type of work, especially as some funders may not understand the richness of the data
- Increasing data costs, especially for primary care data, described by one data user as 'prohibitively expensive', is leading to a lack of diversity of researchers using these data. Only some well-funded research groups can afford the data licences

Facilitators for linked EHR research

Suggestions from data users to facilitate research using linked EHRs in general as well as linked EHRs and infection surveillance data included the following:

Promoting change in practice

- Data provider organisations should be encouraged to recognise that data linkage and sharing is beneficial and should be part of their routine services. There should also be support for a diversity of appropriately qualified and regulated data linkers
- Have mandated data sharing. It was suggested that having an expectation of data sharing early on in the data generation pathway would enhance transparency of processes
- Encourage more reuse of data with really good curation of datasets that could be reused for a broad purpose

Increasing data provider capacity for linked data work

- Funding should be identified for data providers and linkers to embed research within their organisations and develop their services and staff capacity e.g. to fulfil linkage requests from external researchers in a timely manner
- Have systems to place researchers with honorary contracts within data providers
- Have more trusted third parties to do data linkage i.e. any organisation that meets standards of data security and governance. This diversification of data processors and data linkage environments would enable some data linkers to develop areas of specialist knowledge
- Continually develop more innovative approaches to developing good quality data linkages
- Develop connections between PHE and other data providers to establish new linked EHR-infection surveillance datasets

Increasing transparency and communication

- Have a catalogue of national datasets available for research
- Develop a national showcase of examples of the research that people have done with the data
- Enable access to other people's applications for approvals
- Monitor data providers' use and sharing of data to encourage accountability
- Encourage more engagement and communication between data providers or linkers and researchers to enable feedback loops whereby if a researcher discovers a mistake it can be fed back to the data provider and corrected
- Share and publish algorithms, codes, software tools and methods.
- Have more presentation of and discussion about methodology e.g. at the Farr Institute conferences

Streamline data access

- Data providers should be more transparent about the requirements, processes and timescales for accessing data (e.g. like the Scottish model). A roadmap of the process and collaboration requirements for particular data providers and datasets would be useful
- Have a more streamlined data access and approvals process, covering ethics, confidentiality advisory group and R&D, for example to remove red tape
- Have one portal to access data, e.g. at CPRD, data from general practices with different software systems can be accessed through the same portal
- Have a single national information centre across the whole of the UK that holds data very securely and is able to link different datasets in a confidential manner and release de-identified linked data to approved researchers for approved projects

Better support for researchers

- Provide training on the tools and methods required to work with these complex datasets
- Improve support at university research offices for linked database work
- Ensure that researchers and funders recognise the time, funding and administrative requirements to setting up new data linkages

Data users noted that the appointment of a new director of the UK Institute for Health and Biomedical Informatics Research and the existence of HPRUs should enable some new linkages to be effected and for linked infection surveillance data to be used more widely for research.

Examples of good practice

Data users suggested a range of examples of good practice in data linkage and sharing for research with reasons for their choices. This list is not intended to be exhaustive but highlights some key features of successful systems. In general it was noted that data linkage and sharing works well in places where data users and data providers are part of the same organisation. However, while this tends to encourage policy-relevant research, it may discourage 'blue sky' thinking.

Within the UK

Scotland

Reasons given:

- Ability to link multiple routine datasets
- Smoother, more streamlined and less time-consuming data access processes
- Service offered by data providers provides value for money
- Data providers, including health protection teams, keen to work with external researchers

Limitations noted:

- Small population size
- Geographical variation in data completeness

See case study on p17

Wales

Reasons given:

- The Secure Anonymised Information Linkage (SAIL) databank contains many different datasets and some pre-linked data
- Closer synergy between researchers and data providers
- Streamlined approvals process
- Transparent processes for linkage
- Publications on methods show interest in improving quality and transparency

Limitations noted:

- Small population size

UK Biobank

Reasons given:

- Data contents and potential uses for research are completely clear
- Transparent application process
- Small non-prohibitive fee covers access costs
- Any bona fide research is approved and published online
- Data users are required to publish their findings

CALIBER research platform

Reasons given:

- Contains a variety of linked data e.g. EHR from primary care, coded hospital records, social deprivation information and cause-specific mortality data
- Provides an extra layer above existing data such as algorithms to identify outcomes and additional derived variables

Within Europe

Scandinavian countries

Reasons given:

- In general, they have high quality EHR datasets which are automatically linked
- National identifier facilitates data linkage

Limitations noted:

- Lack of primary care records in Denmark
- Problems obtaining updated data (although this is common across settings)

Switzerland

Reasons given:

- National identifier present
- Public expectation that data are used to improve population health e.g. through measuring vaccine uptake and monitoring outbreaks of infectious diseases

Rest of the world

Brazil

Reasons given:

- Much data linkage work is done e.g. the TB register is being linked to mortality and social care data
- Complete public acceptance for linkage of routine health data for research
- Tremendous political will and leadership

Institute for Evaluative Clinical Sciences, Ontario, Canada

Reasons given:

- Status as a research institution with permission to collect government-owned data
- Excellent data linkage systems including for laboratory infection surveillance data linked to EHRs
- In-house data managers are very experienced at extracting data for research purposes
- Good communication between data managers and data users who are co-located

Western Australia

Reasons given:

- Centralised healthcare systems allow creation of population cohorts
- Feedback loops exist through which researchers can query data quality
- Efficient data linkage system

Limitations noted:

- Lack of transparency over data linkage processes
- Can be long timescales for approvals
- Difficulties obtaining updated data
- Small population size

US Emerge Consortium

Reasons given:

- Innovative approach to data linkage between EHRs from a consortium of hospitals and genomic data allows genomic association studies to be run by pooling data across sites

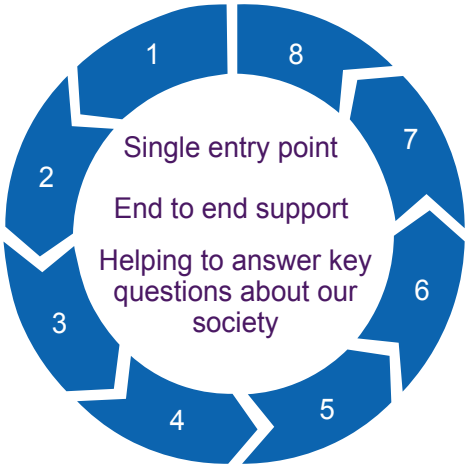
Limitations noted:

- Developing and validating algorithms across all hospital sites is very complex
- Limited data available to external researchers

Case study: Scotland

Several data users described Scotland as a model of good practice for linking and sharing EHR data for research. This may be facilitated by factors such as the smaller population size, more centralised systems and longer-term stability and investment in data infrastructure compared to England. Nevertheless, other features of the Scottish model that facilitate data linkage and sharing include:

- A user-friendly website containing information about all national datasets. Data brochures are also available on request, e.g. Virology and Bacteriology handbooks describing the variables available within infection surveillance datasets
- The electronic Data Research and Innovation Service (eDRIS). This supports the Farr Institute Scotland, the Administrative Data Research Centre and Scottish Government Linkage projects to conduct efficient convenient research. It provides tailored support for researchers to assist with study design, approvals and data access, as per the diagram below:



1. A named research co-ordinator from start to finish
2. Help with study design
3. Expert advice on coding, terminology, meta data and study feasibility
4. Assistance with obtaining the required permissions for access to data
5. Agreed deliverables and timelines
6. Liaisons with data suppliers to secure data
7. Access to data within the NSS National Safe Haven
8. Analyses, interpretation and intelligence about data (where required)

Source: Information Services Division, Scotland.

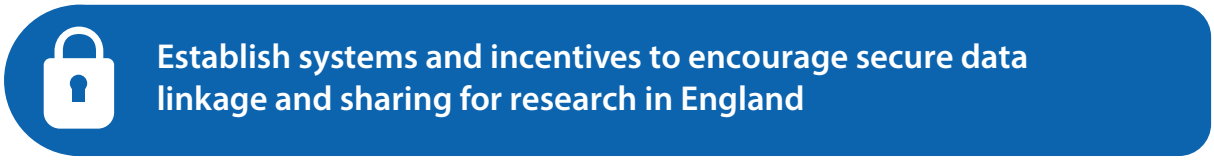
- eDRIS research co-ordinators. Each project is assigned a co-ordinator to help navigate the eDRIS process, discuss data requirements and costs, complete the required approvals, access the NSS National Safe Haven and review and advise on disclosure control for the final project outputs
- A streamlined approvals process. The Public Benefit and Privacy Panel (PBPP) scrutinises all research using NHS-derived data in Scotland, including linked datasets. Researchers typically apply for PBPP approval alongside university research ethics committee approval
- Use of Community Health Index number as an identifier in multiple datasets facilitates data linkage
- Capacity to fulfil data linkage and extraction requests in a timely manner
- Costs that are typically <£10,000 for a project requiring a bespoke linked data extract
- A showcase of examples of research carried out using linked Scottish data

Improving linkage and sharing of routine data for research in England

Despite large investments in big data, we are far from realising the anticipated revolution in personalised healthcare. The interviews highlight the challenges faced by data users in accessing and using de-identified linked electronic health data for research. It is recognised that linking and sharing data with researchers is not the primary function of some data provider organisations (and their views may differ from those captured in this report). Nevertheless, the increasing drive towards digitisation means that it is essential to improve the use of routine health data in England. This is especially vital in these uncertain political times as Brexit begins to exert its effect on Europe-wide plans for data harmonisation and sharing[14].

Ultimately we should aim to maximise the societal benefits of research using routine health data. These include facilitating the conduct of research using routine data to inform both improvements to health and continual refinements to the quality and delivery of healthcare[3,15]. This involves creating structures for secure data linkage and sharing that operate within clear legal and information governance frameworks[16], while being transparent and responsive to research needs. Organisations that will be key to improving systems for routine data sharing and linkage in England include NHS Digital, Public Health England, primary care data providers and research funders.

Recommendations for policymakers concerned with health data



Establish systems and incentives to encourage secure data linkage and sharing for research in England

- Encourage visible national leadership to promote data linkage and sharing e.g. through the UK Institute for Health and Biomedical Informatics Research
- Have mandatory data linkage and sharing standards including considering standards of inter-operability and future proofing for data sharing when setting up a new dataset
- Generate performance metrics to improve quality, transparency and standardisation of data linkage and sharing practices for research
- Encourage public understanding of and feedback on use of these datasets for research through regular consultations

Increase capacity for data linkage and sharing by public organisations

- Ensure adequate capacity for data linkage and sharing within data providers, including developing innovative approaches to data linkage in special populations
- Encourage hosting of researcher placements within all data provider organisations
- Diversify the organisations that are able to act as trusted third parties for data linkage, while ensuring appropriate standards of information security and governance

Streamline procedures to enable appropriate and efficient access to routine health data for research

- Have a national catalogue of datasets collected by each major data provider with notes and a key to understand the variables collected and data quality. This should include datasets not available for external research, with the reason for lack of availability given
- Publish a roadmap for researchers on how to access all routine health datasets, including a contact point for help. An example is shown in the document *Obtaining data from NHS Digital for health research – a guide for researchers* published by the MRC, NHS Health Research Authority and NHS Digital[17]
- Standardise data access and permissions requirements across similar routine health datasets
- Consider allowing re-use of newly generated linked EHR datasets for a broad purpose, as determined by data providers using a risk-based approach

Improve transparency and communication around routine health data access and use between data provider organisations and researchers

- Publish a national showcase of examples of research using different linked EHR datasets
- Have a platform for sharing of codes, algorithms and methods to analyse routine health data
- Encourage development and use of publication standards for research using routine health data, such as the RECORD statement[1]
- Promote dynamic feedback systems between data users and providers to incorporate on new data linkages, feedback and comments on existing datasets and extracts

Provide better support for researchers working with routine health data

- Provide high quality training for data users on research tools, methods, information governance and processes for accessing and using routine health data for research
- Improve the ability of university research offices to support researchers applying for approvals to work with routine health datasets

References

1. Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Med. 2015;12(10):e1001885.
2. Morrato EH, Elias M, Gericke CA. Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in Australia, the UK and the USA. J Public Health Oxf Engl. 2007;29(4):463–71.
3. Gilbert R, Goldstein H, Hemingway H. The market in healthcare data. BMJ. 2015;h5897.
4. Parliamentary Office of Science & Technology. Electronic health records. London, UK: Houses of Parliament; 2016 Feb. Report No.: 519.
5. Parliamentary Office of Science & Technology. Surveillance of Infectious Disease. Houses of Parliament; 2014 Mar. Report No.: 462.
6. National Health Service. NHS Five Year Forward View. 2014 Oct.
7. National Information Board. Personalised health and Care 2020: Using data and technology to transform outcomes for patients and citizens. 2014 Nov.
8. Department of Health. Public Health Surveillance: towards a public health surveillance strategy for England. 2012 Dec.
9. Public Health England. From evidence into action: opportunities to protect and improve the nation's health. 2014 Oct. Report No.: 2014404.
10. Public Health England. Strategic plan for the next four years: better outcomes by 2020. 2016 Apr. Report No.: 2016024.
11. The Review on Antimicrobial Resistance. Tackling drug-resistant infections globally: final report and recommendations. 2016 May.
12. Overview of the General Data Protection Regulation (GDPR). Information Commissioner's Office.
13. COLLIE V COMMON SERVICES AGENCY FOR THE SCOTTISH HEALTH SERVICE. Scotland; May 26, 2010.
14. Auffray C, Balling R, Barroso I, *et al.* Making sense of big data in health research: towards an EU action plan. Genome Medicine 2016; 8:71.
15. Faden RR, Kass NE, Goodman SN, *et al.* An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Ethical Oversight of Learning Health Care Systems. 2013; 43(s1): S16-S27
16. Caldicott F. Information: To share or not to share? The Information Governance Review. 2013 Mar.
17. Medical Research Council, Authority NHR, NHS Digital. Obtaining data from NHS Digital for health research – a guide for researchers. 2017 Mar.

Appendix 1 Interview methods and schedule

Interview methods

Purposive sampling was used to identify academics with experience of using linked electronic health records including infection data for research. Semi-structured interviews were conducted with ten researchers of post-doctoral level and above from several higher education institutions. Interviews were conducted either face to face (n=8) or on the telephone (n=2). All interviews were written down and seven face to face interviews were also recorded, with recordings used to check and provide additional details to the written record. A thematic analysis was carried out to inform this report and the previously published briefing note.

Interview schedule

1. What experience have you had of using electronic health records (EHR) linked to other datasets e.g. infection surveillance data for research?
2. What are the advantages of using linked EHR data in your work?
3. What are your experiences of obtaining access to linked EHR data for research?
4. What are the barriers to conducting research using:
    a. Linked EHR data in general?
    b. Linked EHR and infection surveillance data?
5. What would help to facilitate the effective conduct of research using:
    a. Linked EHR data in general?
    b. Linked EHR and infection surveillance data?
6. Can you give any examples of good practice from elsewhere (national or international) of:
    a. Linking EHR and infection surveillance data?
    b. Sharing linked EHR data between organisations?

Glossary

CPRD – Clinical Practice Research Datalink

eDRIS – Electronic Data Research and Innovation Service

EHR – Electronic Health Record

GDPR – General Data Protection Regulation

HES – Hospital Episode Statistics

HPRU – Health Protection Research Unit

NSS – National Services Scotland

ONS – Office for National Statistics

PBPP – Public Benefit and Privacy Panel

PHE – Public Health England

THIN – The Health Improvement Network

phg foundation

making science work for health

About the PHG Foundation

The PHG Foundation is a pioneering independent think-tank with a special focus on genomics and other emerging health technologies that can provide more accurate and effective personalised medicine. Our mission is to make science work for health. Established in 1997 as the founding UK centre for public health genomics, we are now an acknowledged world leader in the effective and responsible translation and application of genomic technologies for health.

We create robust policy solutions to problems and barriers relating to implementation of science in health services, and provide knowledge, evidence and ideas to stimulate and direct well-informed discussion and debate on the potential and pitfalls of key biomedical developments, and to inform and educate stakeholders. We also provide expert research, analysis, health services planning and consultancy services for governments, health systems, and other non-profit organisations.

978-1-907198-27-4

CAMBRIDGE UNIVERSITY
Health Partners
Knowledge-based healthcare

PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
T +44 (0) 1223 761 900
www.phgfoundation.org

phg foundation
making science work for health