HOUSES OF PARLIAMENT
PARLIAMENTARY OFFICE OF SCIENCE & TECHNOLOGY

POSTNOTE

Number 474 July 2014

Big Data and Public Health

Patient health records and other large scale medical and administrative datasets are increasingly being considered as a valuable tool for the study and improvement of health. This POSTnote examines the sources of data, their current and potential uses for health improvement, and the legal and practical issues raised by data use for public health or research purposes.

Overview

- Large-scale medical and administrative datasets can be used for health service management and research.
- Data sources include prescription data, GP records and education records.
- Medical records may be used in direct patient care, healthcare planning, public health monitoring and academic research and are increasingly being linked to other sources of health and administrative data.
- There are recent and ongoing changes to UK and EU laws relating to the use of patient records beyond direct care.
- Issues with the use of medical records beyond direct care include: public attitudes; timely access to data; privacy, security and identifiability; and data quality and accuracy.

Background

The NHS holds millions of electronic medical records on the health of the population from birth to death. Increased integration and analysis of these alongside other datasets may provide insights that can improve the understanding and management of the population's health. However, the use of personal identifiable data is regulated under the Data Protection Act 1998 and other laws that usually require sensitive information such as individual medical records to be de-identified unless the individual has consented to their use (Box 1).

This POSTnote is part of a series of notes covering the theme of big data. Medical data meet many of the defining parameters for big data such as being large in volume, containing a variety of data formats, and often needing to be accessed quickly (see Big Data Overview, POSTnote 468 for more detail of what big data is).

Box 1. De-identification of patient data

Individual patient records are examples of **identifiable data** because they contain identifiers such as NHS number, name and date of birth from which individuals can be easily identified. They can be subjected to different levels of de-identification. Broadly speaking these include:

- **pseudonymised** (also called key-coded) – identifiers are separated from the record and replaced with a code
- **anonymised** – all personal identifiers are removed
- **aggregated** – multiple records are combined to produce summary level statistics that do not include individual level data.

Sources of Data

There are several sets of data that cover a large proportion of the population and have the potential to be used for population health management. These include NHS records from GPs, hospitals and other settings, along with administrative datasets held by the public sector. Other sources of data that pertain to subsets of the population (cohorts) include genetic information and biobank samples (POSTnote 473), clinical trials data (POSTnote 461) and public survey results gathered across a prolonged period on one cohort. There is also the potential to include data from other sources such as patient experiences shared on social media platforms, and data from supermarkets on consumer habits. The principal sources of data considered within this note are NHS records and administrative datasets collected by government departments and agencies (Box 2).

NHS Records

The conversion of health records into electronic form has provided new opportunities for their use and linkage within and outside the health service for public health

Box 2. Organisations that extract and link health data
There are a number of organisations that routinely extract and link health datasets within 'accredited safe havens' (POSTnote 468).

- The **Health and Social Care Information Centre (HSCIC)** collects Hospital Episode Statistics (HES) and other hospital datasets at a national level and routinely links patient level data from some of these datasets. It plans to collect GP records at a national level as part of the care.data programme (see Box 3). It is an arm's length body of NHS England established in 2013 to replace the NHS Information Centre (IC).
- The **Clinical Practice Research Datalink (CPRD)** contains pseudonymised records from around 8.5% of GP practices which can be accessed securely for health research. It also links these records with other England-wide medical records. It was initially set up in 1987.
- **Clinical Commissioning Groups (CCGs)** employ commercial health informatics specialists to extract and link data from GP and hospital care settings for commissioning purposes at a local level.
- **Public Health England (PHE)** collects and processes vast amounts of data from settings such as GP surgeries, hospitals and NHS laboratories as part of its health surveillance and protection activities. PHE also links its bespoke data collections to existing datasets such as HES and the Office of National Statistics' mortality data. It also generates data via activities such as the genomic sequencing of infectious disease agents.

management and research. The NHS generates a vast range of data including **GP records** of individual patients' illnesses and treatments, and **Hospital Episode Statistics** (HES) containing data from all English NHS hospitals about all attendances, diagnoses and treatments. NHS England plans to link these data under the care.data scheme (Box 3). Additional datasets include prescription records and imaging data such as X-rays and MRI scans.

Administrative Data

Administrative data are data routinely collected by government and other public sector organisations for purposes such as registration, transaction and record keeping. Examples of UK datasets held by government include the National Pupil Database that holds information on school attainment, various indices of deprivation, tax payment records, benefit records and birth and death records. There are currently few examples of research using linked administrative data because of cultural and legislative barriers. A new Administrative Data Research Network aims to enable systematic linkage of data for research purposes.[1] Administrative data about specific individuals can also be linked to their medical records (see Boxes 4 and 5). This enables researchers to study the links between health patterns and factors such as education, environment or socio-economic status at a population level.

Uses of Data

The use of medical records for health improvement purposes can be broadly broken down into primary and secondary uses within the NHS, and secondary uses beyond the NHS for public health and research purposes, regulated by the laws outlined in Box 6. Secondary use of data both within and outside the NHS may require identifiers in order to ensure that linked records refer to the same individual.

Box 3. care.data
In 2013, NHS England commissioned HSCIC to develop care.data, a programme to link individuals' medical records from GP practices and hospitals at a national level.[2] HSCIC already collects patient level data from hospitals and plans to collect patient level GP records through the General Practice Extraction Service (GPES). The aim is to link the datasets for each patient using identifiers such as NHS number or date of birth. Records will be de-identified before any further use. The scheme has raised concerns from the public, GPs, Parliament and the GPES Independent Advisory Group (IAG) such as:

- whether the extent of extraction by GPES might be excessive under the Data Protection and Human Rights Acts
- that the leaflets used to inform people about the scheme were not widely received or read and did not contain sufficient information
- the lack of a straightforward 'opt out' system for patients
- the fact that data would not be pseudonymised at source
- the usefulness of the collected data given that some information would not be extracted and only recent records would be collected
- that personal data might be sold or leaked to commercial organisations such as insurance companies
- the scheme might conflict with GPs' duty of confidence to patients.

Implementation of the scheme has been delayed while these concerns are addressed. The care.data programme board has commissioned an advisory group, made provisions for an 'opt out' for patients and plans to implement a 'secure data lab'. There are concerns that the affair may have affected public attitudes towards health data programmes in general.

Primary Use Within the NHS

Primary use in the case of medical records refers to direct care, as doctors need information about their patients to make decisions about treatment. These data may need to be shared between members of a patient's care team. Sharing of data for direct care purposes usually relies on implied consent from the patient.

Secondary Use Within the NHS

The NHS also makes extensive use of data beyond direct patient care. For instance data informs commissioning, clinical audit, treatment outcome monitoring, calculation of treatment costs and payment to practices. Some of these functions are carried out 'in-house' while others are outsourced to commercial companies. Many functions require linkage of NHS datasets such as GP patient records, HES and prescribing information. Examples of NHS secondary data use include:

- **clinical audit** to assess the standard of care provided by GP practices and hospitals to identify areas which are exceeding or falling short of expected standards.
- **risk stratification** to identify groups or individuals potentially at high risk of disease development or progression to allow timely intervention or treatment.
- **commissioning** of NHS care at a local level via CCGs (Box 2) or at a national level. Patient information is used to identify population needs in order to select the most efficient and effective services and providers.

Secondary Use Beyond the NHS

Secondary uses of medical data beyond the NHS may be broken down into: local and national public health activities; academic research; and data use by commercial

Box 4. National Cancer Intelligence Network Routes to Diagnosis
The National Cancer Intelligence Network (NCIN) Routes to Diagnosis study examines different routes to cancer diagnosis, including delays in diagnosis, and their impacts on survival. It links data from Hospital Episode Statistics, cancer waiting times and cancer screening to data from the National Cancer Data Repository. Personal identifiers are used to link these datasets at patient level and to look at the effects of factors such as socio-economic status, age, gender and ethnicity on Route to Diagnosis and patient outcome, by cancer type. Results have fed into public awareness campaigns such as PHE's Be Clear On Cancer Campaign, with the aim of helping patients to spot symptoms of cancer earlier.

Box 5. Avon Longitudinal Study of Parents & Children (ALSPAC)
ALSPAC is a long term research project charting the health and development of around 14,500 individuals who were born in 1991-1992 in the Bristol area.[3] The study's parents and children provide questionnaire data, clinical data, and biological and genetic samples. These are linked to health and administrative records such as GP records and education data from the National Pupil Database. Additional secure data sharing with government departments is being arranged in conjunction with the Farr Institute and the Administrative Data Research Network, and ALSPAC is developing anonymous record linkage procedures that do not require any data extraction.[4] ALSPAC operates a broad consent model for the use of data, and participants can opt out of specific research projects on a case-by-case basis. ALSPAC has published over 1,000 research papers to date with many findings used to inform UK and international health policy. These include findings that the consumption of oily fish benefits childhood IQ and development and that peanut oil in baby creams may trigger nut allergies.

organisations. But there is not always a clear distinction between these. For example, Public Health England (PHE, see Box 2) commission academic research into public health, and research partnerships often involve academia, charities and industry.

Public Health Monitoring and Management
Medical and administrative records are used to carry out public health monitoring and management. National and local authorities conduct surveillance of infectious diseases and environmental hazards for public health protection purposes, along with the monitoring of non-infectious diseases such as cancer, with a view to improving treatment efficiency and outcomes (Box 4).

Academic Research
The linkage of medical records to cohort studies and trials has enabled research into population health such as that linking smoking to lung cancer.[5] Linkage across different healthcare datasets is also of value in identifying new risk factors and highlighting novel pathways for treatment. Multiple ongoing large-scale long-term cohort studies collect and link datasets to improve understanding of population health (Box 5 and POSTnote 473). There has been recent investment in a number of research institutes specialising in the use and linkage of large medical datasets, such as the FARR institutes.

Use by Commercial Organisations
The pharmaceutical industry uses medical data to monitor drug safety (required by statute) and efficacy at a population level. Findings are fed back to clinicians to improve the efficacy of care pathways. The industry is seeking to make greater use of health records and other data to improve drug development and provide a targeted approach to medicine which uses a patient's genetic, health and lifestyle data to inform treatment decisions (stratified medicine).[6] Other commercial organisations, including insurance companies, have previously accessed and used pseudonymysed HSCIC datasets. Access to these data is now only permitted for healthcare provision or promotion (Box 6).

Box 6. Legislation of access to medical records
In addition to the Common Law Duty of Confidentiality, several pieces of primary and secondary legislation apply to use of medical records.

- **The Human Rights Act 1998** sets out the right to privacy and a family life, with no interference from the state except for specific lawful purposes such as health protection.
- **The Data Protection Act 1998 (DPA)** balances an individual's rights to privacy with the requirement of organisations to collect and use personal information. Under the Act, medical records are classed as personal sensitive information subject to stricter access requirements than other personal data. The Act sets out a duty of fair processing that requires data controllers to inform data subjects of how their information is being used and requires that data use is accurate and not excessive. Anonymised data are no longer classed as personal data.
- **The Health Service (Control of Patient Information) Regulations 2002** provide a statutory gateway for the collection of confidential patient information relating to neoplasia (abnormal cell growth including cancer) and infectious disease.
- **Section 251 of the National Health Service Act 2006** makes provisions for the use of identifiable records without the consent of the data subject, where obtaining consent is not feasible and use of data is in the interests of the patient or of the wider public.
- **The Health Protection (Notification) Regulations 2010** place a duty on healthcare providers to notify the Health Protection Agency of incidences of infectious diseases.
- **The Health and Social Care Act 2012** provides a statutory gateway for the collection and processing of confidential personal data by the Health and Social Care Information Centre (HSCIC).
- **The Care Act 2014** amends the Health and Social Care Act to prevent HSCIC disseminating data unless it is for the provision of health and social care or the promotion of health.

Challenges Related to Data Use

General issues concerning big data are discussed in POSTnote 468. The following sections examine technical issues specific to the collection and use of health-related data, including data extraction and linkage, and data quality, along with the governance challenges of maintaining data security and public support while allowing data access for public benefit.

Technical Challenges

Data Extraction and Linkage
Linkage of an individual's data from multiple care settings can give a more complete picture of their health status and treatment pathway. Because there is no central database holding all such records, datasets must first be extracted (collected) from their original source (such as GP computer systems) before they can be linked. Currently there is no national standard mechanism for medical data extraction and processing. The type and quality of data extracted, the method and frequency of extraction and the extent of

linkage to other datasets depend on the sources of data used and the intended purpose of the programme. The care.data programme (Box 3) is an attempt to link data from multiple settings at a national level. However, it also highlights some of the problems that can be encountered when attempting to link large sets of medical data.

Data Quality and Accuracy

There are concerns about the quality, completeness and accuracy of health-related data. Clinicians are increasingly using pre-assigned codes to record illnesses and treatments, rather than free text, which leads to the possibility of incorrect codes being used. Further issues include missing information such as cessation of medicine use, and duplication and invalidation of NHS numbers and other identifiers that can affect data quality and linkage.

Legislation and Governance Challenges

Public trust in the governance of data use is considered key to the continued and expanded use of medical records for health research and management. Surveys suggest that the public is broadly supportive of the use of data for medical research. However, numerous concerns remain. These include: use without consent; use of identifiable data; data security; lack of transparency; potential discrimination by employers or insurers; and access by commercial organisations.[7, 8, 9] Such concerns may result in patients withholding information from healthcare providers, which may be detrimental to the patient and reduce data quality.

UK Legislation Regulating Data Access

Healthcare providers have a duty of confidentiality to their patients and must seek a patient's consent before sharing his or her personal data. Identifiable medical records are also classed as sensitive data under the DPA and their use is therefore strictly regulated. As outlined in Box 6 there are several statutory provisions for the use of this data without consent, such as monitoring of neoplasia (cancer) and infectious disease. Approval for other uses of identifiable health data without explicit consent can be granted under section 251 of the NHS Act (Box 6). However, some section 251 approvals, such as those for routine healthcare planning, have been criticised by privacy advocates, who argue that the measure was intended as an extraordinary provision for use in high priority public health work and research. There are also calls to make more use of explicit consent models for secondary uses of data rather than relying on implied consent or statutory provisions.[10] Explicit consent is always required for the use of identifiable data from social care settings. The new Care Act aims to restrict data release from the HSCIC to that for appropriate health related purposes (Box 6). However, concerns remain about the breadth of access that might be permitted on 'health promotion' grounds.

The European Data Protection Regulation

A new draft European Data Protection Regulation is currently being debated. There is uncertainty about how the European Parliament's amendments to the text will impact on the use of personal identifiable data for public health and research purposes.[11] In particular, there are concerns that the consent requirements for use of identifiable data in research laid out in article 81 of the current draft may be incompatible with the broad consent models used in many studies. Individual member states may make provisions for de-identified data to be used without consent under certain circumstances, but there is no clear provision for the use of identifiable data without consent, such as that granted under section 251 of the NHS Act. It is not clear how the proposed regulation would affect public health activities. There are provisions for the retention and use of data for health related purposes, but concerns remain about the possible impact on public health monitoring activities, such as those in Box 4.[12]

Governance of Data Access

A recent review has highlighted lax governance of data release by the HSCIC's predecessor, the NHS IC.[13] Despite this, researchers reported significant administrative burdens in gaining access to medical records, such as the need for approval from multiple advisory bodies, leading to campaigns for improved access for research.[7,14] Lengthy access procedures conflict with the need for timely access to data needed for research addressing immediate health concerns and/or conducted on short-term grants. The HSCIC is currently reviewing its data release procedures to address the failings of its predecessor. However, public health officials report that this is delaying their access to data for important disease surveillance activities. Increased extraction and use of medical records also increases the burden of fair processing placed on GPs under the DPA to ensure that patients are aware of how their data are used.

Data Security and Identifiability

Security experts have demonstrated that there is always a risk that patient level data could be re-identified even if it has been anonymised or pseudonymised.[15] The linkage of multiple records about one individual may increase data usefulness but requires the use of identifiers to do so, and may increase the risk of re-identification. This is both a security and a legal issue since the use of identifiable data without consent is limited by law (Box 6). Data security may be increased by use of advanced data linkage technologies that do not require extraction of identifiable data.[4, 16]

Endnotes

[1] ESRC Press Release on ADRN, 2013
[2] House of Commons Standard Note: Care.data, 2014
[3] Boyd *et al., International Journal of Epidemiology*, 1–17, 2012
[4] Wolfson *et al. International Journal of Epidemiology,* 39, 1372–1382, 2010
[5] Doll *et al.* Br J Cancer, 92(3), 426–429, 2005
[6] ABPI, Big Data road map, 2013
[7] BHF, Policy Statement: Patient data in research and healthcare, March 2012
[8] Wellcome Trust, Qualitative Research into Public Attitudes, 2013
[9] Ipsos MORI, Dialogue on Data, 2014
[10] Robin Burgess, London Connect Blog: Why consent matters so much, 2013
[11] Protecting health and scientific research In the Data Protection Regulations (2012/0011(COD), Wellcome Trust, 2014
[12] EUPHA Factsheet: Revision of the European Data Protection Legislation, 2013
[13] Partiridge, N., HSCIC Data release review, 2014
[14] AMRC Statement of the use of patient data for research, 2013
[15] Ohm P, 'Broken promises of privacy', *UCLA Law Review,* 57, 2010.
[16] The ONS Virtual laboratory, accessed 04/07/2014

POST is an office of both Houses of Parliament, charged with providing independent and balanced analysis of policy issues that have a basis in science and technology. POST is grateful to Helen Brewer for researching this briefing, to the BBSRC for funding her parliamentary fellowship, and to all contributors and reviewers. For further information on this subject, please contact the co-author, Pete Border. Parliamentary Copyright 2014. Image copyright iStockphoto.com